ISSN: 2278 – 7798
International Journal of Science, Engineering and Technology Research (IJSETR)
Volume 5, Issue 5, May 2016

Big Data Analytics and Cloud Computing: Parallel Processing of Massive Data Sets to Accelerate Performance

Siddhant Wadhwani[1], Kaavya Wadhwani[2], Geocey Shejy[3]

[1]*Department of MCA, V.E.S.I.T., Mumbai University*
Mumbai, Maharashtra, India

[2]*Department of MCA, V.E.S.I.T., Mumbai University*
Mumbai, Maharashtra, India

[3]*Department of MCA, V.E.S.I.T., Mumbai University*
Mumbai, Maharashtra, India

Abstract- **Data has been increasing manifold due to large number of companies established in recent years, development of social networks, logs and historical records collected and also cloud computing. This has led to a new challenge for Extraction, Transformation and Loading (ETL process) of data, which means storing, processing, and analyzing large volumes of data. The traditional technologies are not a perfect solution to process Big Data. Big Data platforms help users to develop analysis services effectively. However, it still takes a long time to collect data, develop algorithms and analytics services.**

Companies providing cloud-scale services have an increasing need to store and analyze massive data sets such as search logs and click streams. For cost and performance reasons, processing is typically done on large clusters of shared-nothing commodity machines.

It is imperative to develop a programming model that hides the complexity of the underlying system but provides flexibility by allowing users to extend functionality to meet a variety of requirements.

In this research paper, we present a new declarative and extensible scripting language and programming model, which is targeted for massive data analysis.

The language is designed for ease of use with no explicit parallelism, while being amenable to efficient parallel execution on large clusters. It uses a combination of languages that include C#, SQL and also tools integrated within the model for providing analysis services.

Keywords- **Big Data, Cloud, Hadoop, MapReduce, Cloud Computing**

I. INTRODUCTION

Recently, a great deal of interest in the field of Big Data and its analysis has risen, mainly driven from extensive number of researchchallenges strappingly related to bonafide applications, such as modeling, processing, querying, mining, and distributing large-scale repositories. Big Data is a term that describes the large volume of data – both structured and unstructured, that inundates a business on a day-to-day basis. It's not the amount of data that's important but what organizations do with the data that matters. Big Data can be analysed for insights that lead to better decisions and strategic business moves. The amount of data that's being created and stored on a global level is almost inconceivable, and it just keeps growing. That means there's even more potential to glean key insights from business information – yet only a small percentage of data is actually analysed.

Big Data analysis is somehow a more challenging task than locating, identifying, understanding, and citing data. By analyzing simple data having one data set, a mechanism is required to design a database. There might be alternative ways to store all of the same information.

In Big Data, data is rather a "fuel" that powers complex of technical facilities and infrastructure components built around a specific data origin and target use. There are not many academic papers related to Big Data; in most cases they are focused on some component technology (e.g. Data Analytics or Machine Learning) or solutions that reflect only a small part of the whole problem area. The same relates to the Big Data definition that would provide a conceptual basis for further technology development. There is no well-established terminology in this area. The importance of Big Data doesn't revolve around how much data you have, but what you do with it.

You can take data from any source and analyze it to find answers that enable 1) cost reductions, 2) time reductions, 3) new product development and optimized offerings, and 4) smart decision making.

Big Data combined with high-powered analytics can help accomplish tasks such as: 1) Determining root causes of failures, issues and defects in near-real time,
2) Generating coupons at the point of sale based on the customer's buying habits, 3) Recalculating entire risk portfolios in minutes, and 4) Detecting fraudulent behaviour before it affects your organization.

Big data affects organizations across practically every industry. The main motive of many companies is 'To Simplify and Accelerate with Platform as a Service (PaaS)'. The various cloud services offered are Infrastructure as a Service (IaaS), Platform as a Service (PaaS), Software as a Service (SaaS) and Data as a Service (DaaS). Main strategies depend on integration, process, mobile, analytics, collaboration, security and data management as platform for change. Most employees, project leads, senior software managers, DBAs, project managers, POCs, etc. have no clue as to how should one migrate their on-premise databases to the cloud. They lack the knowledge about the process of cloning their repositories and creating backups onto the cloud. While some are confused whether to go with private, public or hybrid clouds, others are busy thinking of the technologies to be used for the same like NoSQL, Hadoop, Oracle 12c, etc.

A new development paradigm is to build cloud native applications following the layered approach. The DevOps model like Ansible, Puppet Labs, SaltStack, Chef, etc. that basically operate by adding new features and keeping the current system stable and fast. Being in the digital world, we have experienced many Digital Disruption Waves from ride sharing to car sharing to iCar (2D to 3D), drones and many others. The future that one can foresee from a technology standpoint, must follow the Agile methodology. During our internship, we too followed the same development methodology, best practices and coding standards to deliver prototypes daily to the clients.

As our current project has been moved to Cloud based on the client's requirements, not a major part of our system architecture had to be modified. The system follows a 3-tier architecture namely: Backend, Middleware and Frontend. Here, the most of the backend part has been moved to cloud. Data that is collected from upstream sources is collected into a staging database where ETL processes are triggered. All execution process has been moved to Cloud using script languages with combination of Big Data, Hadoop, C# and SQL with Integration Services. The middleware layer comprises of the Web API reference service and is connected to the backend on one hand and frontend on the other. The frontend consists of reports and analysis that has been processed and obtained from the backend, represented in the form of reports and visualizations. For this, we use SharePoint Online to create web parts and for the Reporting UI with technologies such as HTML5, CSS3, JavaScript, JSON, AJAX, GULP, Ruby on Rails, etc. to generate Web reports. Also, a windows application is available for generating and viewing the same reports. Microsoft PowerBI and Oracle JET provide appropriate reporting and visualization UI/UX charts. All the configurations information is stored in a Config database and the reports information in a Reporting database.

An alternative to RMS is the Octopus Deploy which helps in easy integration with PowerBI, faster deployment using PowerShell scripts, partially open source technology and being highly cost-effective.

These modern techniques over the cloud can help perform parallel processing for enormous data sets and thus improve the overall performance of our systems. Also, programming languages such as R and SAS are being implemented across many organizations for data analytics.

II. EMERGING TECHNOLOGIES

Azure Data Lake is another innovative technology being implemented by most organizations across the globe. It is a batch, real time system that has made interactive analysis easy. It can be used to store and analyze any kind and size of data. One can interactively explore patterns in data, develop faster, debug and optimize smarter. There is no learning curve and it dynamically scales to match the business priorities. It is built on Hadoop 2.0 YARN (Yet Another Resource Negotiator), specifically designed for the cloud. It consists of the HDInsights, Analysis and Storage components. Hive, Storm, etc. are tools used in Visual Studio. It is based on U-SQL language.

For developing Big Data Applications, it is a must to author, debug and optimize them. One can refer to multiple languages: U-SQL, Hive and Pig. With U-SQL language, we can seamlessly integrate with the existing code in Visual Studio. Some benefits provided by this are: Simplified Management and Administration, Web-based management in Azure portal, automating PowerShell scripts and monitoring services.

Data appears in three native forms: Unstructured, Semi Structured and Structured. Data is normally in *unstructured* format and requires to be *structured* when pushed to Cloud. The concepts of Ingress and Egress appear when data is moved from On-premise to Cloud and vice-versa. Big Data brings value to the public by providing Ease of Use, Tooling, Scalability, Streaming, Machine Learning, Performance, Interactivity and Language options.

Map-Reduce programming model provides a good abstraction of group-by-aggregation operations over a cluster of machines. The programmer provides a map function that performs grouping and a reduce function that performs aggregation. The underlying run-time system achieves parallelism by partitioning the data and processing different partitions concurrently using multiple machines.

However, this model has its own set of limitations. Users are forced to map their applications to the map-reduce model in order to achieve parallelism.

For some applications this mapping is very unnatural. Users have to provide implementations for the map and reduce functions, even for simple operations like projection and selection. Such custom code is error-prone and hardly reusable.

Moreover, for complex applications that require multiple stages of map-reduce, there are often many valid evaluation strategies and execution orders. Having users implement (potentially multiple) map and reduce functions is equivalent to asking users specify physical execution plans directly in database systems. The user plans may be suboptimal and lead to performance degradation by orders of magnitude.

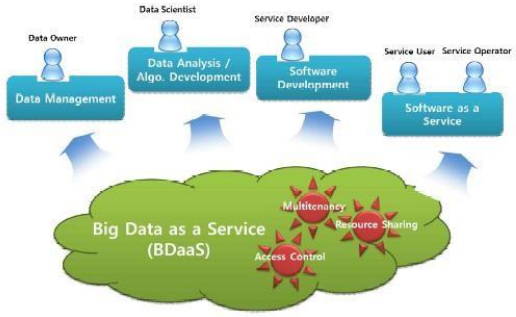

Fig. 1. Collaborative Big Data Platform concept for Big Data as a Service

Most companies offer services to develop and host custom applications with low cost strategy and high availability. The companies migrate existing applications to Azure, enable transition of existing cloud applications to Azure, upgrade and maintain applications on Azure. They optimize cloud solutions for faster access, thereby enhancing user experience and ensuring secure applications comply with latest industry standards.

Most MNCs like Google, Oracle, Microsoft and other top companies offer best-in-class services across software as a service (SaaS), platform as a service (PaaS), and infrastructure as a service (IaaS). Cloud helps organizations drive innovation and business transformation by increasing business agility, lowering costs, and reducing IT complexity. Oracle's Cloud platform, Google's app engine for Cloud, Microsoft and Amazon web services are available for company employees in the IT industry as well as customers and owners across the world as free and paid services.

The three latest emerging cloud technologies paving the way for success in future are – Containerization, Desired State Configuration and Micro services. Containerization takes virtualization to the next level. Desired state configuration (DSC), a new management platform in Windows PowerShell, enables IT to deploy and manage configuration data for software services and manage the environment in which these services run. With micro services, instead of using one chunk of code, developers can build granular services from the bottom up without being tied to a previously rigid infrastructure. Examples of micro services are protocol gateways, user profiles, shopping carts, inventory processing, queues and caches.

These cloud technologies are changing the way IT practices server management.

III. CLOUD COMPUTING

Definition: A style of computing based on shared, elastic resources delivered to users in a self-service, metered manner using web technologies. To fully understand cloud computing across an enterprise, you need to understand the different functional benefits driving cloud's popularity. We need to think about how to craft a cohesive cloud strategy that works for everyone in the enterprise.

A.Cloud Platform

Private cloud deployment is booming, as companies see the benefits of faster provisioning, on demand access, and scalability. Cloud builders can help their IT department become agile, cost-efficient private cloud providers. They view cloud from a lifecycle management perspective, from resource management and monitoring to capacity planning and chargeback mechanisms.

B.Cloud Applications

Accelerate time to value with a complete and integrated portfolio of business application that deploy quickly and respond to changing business requirements. Offload IT management and focus on innovation.

C.Cloud Infrastructure

Complete offering of servers, storage, networking fabric, virtualization software, operating systems, and management software to help customers build a custom-fit enterprise cloud.

D.Cloud Lifecycle Management

Self-service provisioning, centralized resource management, integrated capacity planning, and complete visibility from applications to disk.

E.Cloud Integration

Our modern and open service-oriented platform and infrastructure simplifies integration needs and lowers total cost of ownership.

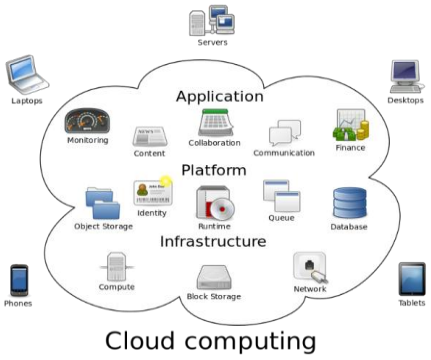

Fig. 2 Architecture of Cloud Computing

The network elements representing the provider-rendered services are invisible, as if obscured by a cloud. Cloud computing, also called On-demand computing, is a kind of Internet-based computing that provides shared processing resources and data to computers and other devices on demand.

It is a model for enabling ubiquitous, on-demand access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications and services), which can be rapidly provisioned and released with minimal management effort. Cloud computing and storage solutions provide users and enterprises with various capabilities to store and process their data in third-party data centres. It relies on sharing of resources to achieve coherence and economy of scale, similar to a utility (like the electricity grid) over a network. Cloud computing has become a highly demanded service or utility due to the advantages of high computing power, cheap cost of services, high performance, scalability, accessibility as well as availability. Some cloud vendors are experiencing growth rates of 50% per year.

IV. WHY MOVE DATA TO CLOUD?

Cloud storage is cheaper, expands endlessly and needs little attention; but how much data can a company realistically park in the cloud? It's a popular belief that moving data to the cloud not only takes advantage of the low prices cloud storage vendors can achieve due to their economies of scale, but it can also free a company from the drudgery of buying, commissioning, provisioning and maintaining storage systems. In addition, if cloud storage is endlessly elastic, it can be used without the kind of careful capacity planning that in-house storage requires.

Any move to the cloudmust first consider the impact such a move would have on your users and their applications. In many cases, it may be feasible to move much of your data to the cloud, but many key applications are likely to require their data to be kept in-house. Still, if the data existed in the cloud, local caches and copies of data can be retained and used as needed [6].

The best way to get files into the cloud, while still making them accessible to users in other locations, is to install an integrated cloud storage appliancein each location. These cloud storage gateways, from vendors such as Amazon Web Services, Avere Systems, Nasuni, Panzura and TwinStrata, use local storage, which can include solid-state drives, as a cache and present the data via Server Message Block (SMB) and/or NFS so users can access their data just as if they had a local NAS while the authoritative copy of thedata is stored in the cloud.

Most of these solutions also use the cloud to store an essentially unlimited number of snapshots. Between these snapshots and the cloud storage provider replicating the data to multiple locations, traditional backups could become a thing of the past.

The biggest challenge is finding a solution that can serve database applications. You could run Microsoft SQL Server or Oracle in an Amazon Elastic Compute Cloud (EC2) instance accessing Amazon's Elastic Block Store (EBS), but adding 20ms to 200ms of latency between the application running on a user's PC and the database server will likely have a negative impact on performance and user experience.

V. SHOULD 100% DATA BE MOVED?

While moving all of a company's data into cloud storage servicesmay be somewhat impractical at this time, there are tools available today to move at least a copy of all your data to the cloud.

Some applications, like email, can be shifted to the cloud fairly easily, while other applications can be replaced with cloud-based apps that offer equivalent functionality. Both of these approaches will move data off premisesand into the cloud. For other applications, especially those that might suffer from the latency that cloud storage incurs, a hybrid approachwhere data is stored locally for performance, but eventually ends up in the cloud, may be best.

VI. DIFFICULTIES IN MOVING DATA

In the case of software as a service (SaaS) [9], however, there could be bigger problems. If the service being offered was based on a standard application – for example, SugarCRM[13]or OpenERP[7]– it should be possible to find another service provider hosting the same application. There may be differences in the implementation, but all that should be required is an extract/transformation/load (ETL)action to make sure the data fits the schema of the implementation, the new service provider has in place[14].

IT executives should remember that any modifications they were allowed to make to the application by the previous provider (such as skinning the application with a logo or the addition of any extra functions) will need to be carried out again with the new provider.

In many cases, it will not be possible to pull any of the changes from the previous provider, so re-implementing these will be the hardest part of the transition. What it does mean is that any changes that are carried out, even in a SaaS environment, must be documented and stored outside of the

SaaS environment – a full change log is necessary so that the changes can be re-implemented if a change of provider is needed.

The real problems come when a business is moving from a provider which has proprietary software in place. This may be a provider which has so heavily modified an open source application as to essentially make it a new application. Or, it could be a SaaS provider which owns the application and does not allow it to be offered by any other cloud provider on their own platform, such as Salesforce.com[15].

Next is the need to identify the schemas used by both systems. Matching field names and types is necessary here to make sure that fidelity of information is maintained when the data is moved across. This will also define the ETL activitythat will have to be carried out. [11]

Then, there is the necessity of testing. It cannot be left to chance and hope that such an activity will work. You will need to carry out a test by taking data from the existing environment and moving it to the new environment. This does not have to be based on a permanent contract with the second provider – it is just a test to make sure it works.

Based on the test being successful, you can create a full, formalised plan as to what your organisation needs to do should the worst come about.

This should also include indications of how long such activities are expected to take – and the plans around how the business will continue to operate during this downtime. This may well involve falling back on manual processes – and any data that is gathered during these manual processes will need to be input into the new system as it comes on line.

The last area that should be covered by the contract is that the old service provider must securely wipe your organisation's data from their systems – something that is overlooked more often than not.

VII. MOVING DATA TO CLOUD

Data movement will become easier as cloud matures. Hopefully, as time progresses and cloud standards bed down, it will be possible to move applications and data about on standardised cloud platforms, such as OpenStack [16], and the whole activity should be a great deal more seamless and easy. Cloud providers are no different to any other commercial entity. There will be failures along the way, and this is no reflection on cloud as a model for implementing an IT platform. The problem is that any failure of a cloud provider will hit more organisations, as they are, by definition, multi-tenanted platforms [17].

IT must be prepared with a strategy to minimise the impact of the failure of their cloud or datacentre provider – whether it be due to a breakdown in relationship or the complete failure of the cloud provider.

VIII. TRENDS IN DATA ANALYTICS

The different trends in data can be analysed based on the following factors:

A.Big Data Analytics in the cloud

Hadoop, a framework and set of tools for processing very large data sets, was originally designed to work on clusters of physical machines. That has changed. "Now an increasing number of technologies are available for processing data in the cloud," says Brian Hopkins, an analyst at Forrester Research. Examples include Amazon's Redshift hosted BI data warehouse, Google's BigQuery data analytics service, IBM's Bluemix cloud platform and Amazon's Kinesis data processing service. "The future state of big data will be a hybrid of on premises and cloud," he says.

B.Hadoop: The new enterprise data OS

Distributed analytic frameworks, such as MapReduce, are evolving into distributed resource managers that are gradually turning Hadoop into a general-purpose data operating system. The ability to run many different kinds of [queries and data operations] against data in Hadoop will make it a low-cost, general-purpose place to put data that you want to be able to analyse.

C.Big data lakes

Traditional database theory dictates that you design the data set before entering any data. A data lake, also called an enterprise data lake or enterprise data hub, turns that model on its head.

D.More predictive analytics

With big data, analysts have not only more data to work with, but also the processing power to handle large numbers of records with many attributes.

E.SQL on Hadoop: Faster, better

Tools that support SQL-like querying let business users who already understand SQL, apply similar techniques to that data. SQL on Hadoop "opens the door to Hadoop in the enterprise," Apache Hive has offered a structured, SQL-like query language for Hadoop for some time.

F.More, better NoSQL

Alternatives to traditional SQL-based relational databases, called NoSQL (short for "Not Only SQL") databases, are rapidly gaining popularity as tools for use in specific kinds of analytic applications, and that momentum will continue to grow. A NoSQL product with graph database capability, such as ArangoDB, offers a faster, more direct way to analyze the network of relationships between customers or salespeople than does a relational database.

G.Deep Learning

A set of machine-learning techniques based on neural networking, is still evolving but shows great potential for solving business problems. It enables computers to recognize items of interest in large quantities of unstructured and binary data, and to deduce relationships without needing specific models or programming instructions.

H.In-memory analytics

The use of in-memory databases to speed up analytic processing is increasingly popular and highly beneficial in the right setting. Bringing in an in-memory database means there's another product to manage, secure, and figure out how to integrate and scale.

IX. HOW CAN COMPANIES MAKE THE MOST OF BIG DATA?

Big data technology is allowing more companies to collect vast amounts of information about their customers. But figuring out what to do with that data while protecting clients' privacy is becoming a big issue for many businesses. Thus, it is mandatory for companies to understand their goals before investing directly. They are directly starting the process of collecting data without understanding what needs to be done and how it can be achieved. We suggest for these circumstances, there are various events held by top companies such as Google, Oracle, etc. where they give in-depth knowledge of how migrating to cloud can benefit and the process can also be explained deeply with more clarity.

Recently, we signed up for an event – Oracle Cloud World Developer, 2016 in Mumbai and got to know about all the practical benefits of Big Data and Cloud, which can help in determining patterns of consumer behaviours that helps in the analysis and thus companies can adapt to the changing markets. Many businesses are already using advanced software to extract relevant data automatically. The opportunity for big data in the world of global commerce is enormous and powerful but remains elusive.

X. ADVANTAGES AND DISADVANTAGES

Storing data on cloud has several pros and cons, but as per our research and analysis, the number of advantages outweigh the disadvantages by a huge margin. The following are a list of advantages and disadvantages based on certain factors:

A.Usability

All cloud storage services reviewed in this topic have desktop folders for Mac's and PC's. This allows users to drag and drop files between the cloud storage and their local storage.

Be careful when using drag/drop to move a document into the cloud storage folder. This will permanently move your document from its original folder to the cloud storage location. Do a copy and paste instead of drag/drop if you want to retain the document's original location in addition to moving a copy onto the cloud storage folder.

B.Bandwidth

You can avoid emailing files to individuals and instead send a web link to recipients through your email.

Several cloud storage services have a specific bandwidth allowance. If an organization surpasses the given allowance, the additional charges could be significant. However, some providers allow unlimited bandwidth. This is a factor that companies should consider when looking at a cloud storage provider.

C.Accessibility

Stored files can be accessed from anywhere via Internet connection.

If you have no internet connection, you have no access to your data.

D.Disaster Recovery

It is highly recommended that businesses have an emergency backup plan ready in the case of an emergency. Cloud storage can be used as a back-up plan by businesses by providing a second copy of important files. These files are stored at a remote location and can be accessed through an internet connection.

E.Cost Savings

Businesses and organizations can often reduce annual operating costs by using cloud storage; cloud storage costs about 3 cents per gigabyte to store data internally. Users can see additional cost savings because it does not require internal power to store information remotely.

F.Data Security

There are concerns with the safety and privacy of important data stored remotely. The possibility of private data commingling with other organizations makes some businesses uneasy.

G.Software

If you want to be able to manipulate your files locally through multiple devices, you'll need to download the service on all devices.

In this section, we understand the benefits of moving data on cloud and can enlist the reasons for the same: cost-effective, invisibility, security, automation, accessibility, syncing, collaboration, protection and data recovery. Cloud storage through an online backup service provides a myriad of advantages for storing both your home and professional files. Thus, data can easily be moved to cloud as - no server maintenance is required, no data loss, more secure and easy to share with clients in real time, to get instant feedback.

XI. COMPARISON OF DIFFERENT SURVEYED TECHNOLOGIES

Based on our research, we analysed various Cloud and Big Data technologies and thus prepared a rough comparison that could help others to choose their technologies and frameworks based on their needs. R and SAS are programming languages that are especially used for analytics. SAS vs R has probably been the biggest debate analytics industry might have witnessed. SAS has been the undisputed market leader in commercial analytics space. The software offers huge array of statistical functions, has good GUI (Enterprise Guide & Miner) for people to learn quickly and provides awesome technical support. However, it ends up being the most expensive option and is not always enriched with latest statistical functions.

Parameters	R	SAS	Preferred
Availability / Cost	5	3	R
Ease of Learning	3	4.5	SAS
Data Handling Capabilities	4.5	4	R
Graphical Capabilities	4.5	4	R
Advancements in Tool	4.5	4	R
Job Scenario	3.5	4.5	SAS
Customer Service Support & Community	3.5	4.5	SAS

Fig. 3 Comparison of R and SAS [Scale: (1 - Low; 5 - High)]

R is the Open source counterpart of SAS, which has traditionally been used in academics and research. Because of its open source nature, latest techniques get released quickly. There is a lot of documentation available over the internet and it is a very cost-effective option.

Of course while SAS and R commands solve many of the same problems, they are certainly not perfectly equivalent.

XII. DESIGN, INNOVATION AND EVALUATION OF OUR PROPOSED SYSTEM

The system that we have invented works on getting data from upstream sources, perform ETL activities on the data which then connects to the Web API (middleware) and then to the front-end. For the backend, we are implementing a new language that includes the combination of SQL and C# that helps to push data onto the cloud in a more convenient manner. It makes use of C# libraries and SQL's in-built functions and combines their productive needs to achieve parallelism for processing data on cloud. Techniques such as Ingress and Egress play an important role in this technique. This technology will be revealed soon and shall be adopted by many other organizations based on their needs and choices.

We have first performed a Clean-up activity which uses a variety of tools like F12 toolkit, SQL search and Fiddler. Then we verify the changes and check if it would affect the performance or not. Various ETLs are run to ensure that we get accurate results. They can be scheduled on a daily or weekly basis. All the business logic is taken care of and a backup of backend database is taken to ensure database recovery and safety. Now, for pushing the updated databases with lesser consumed memory and faster performance, we use a combo of SQL and C# along with PowerShell scripts to configure data to move on Cloud. Tools like SQLtoSS help in converting the unstructured data tables into structured streams on the cloud which helps in parallelizing data and improves the performance. Accurate data is seen on cloud after verifying it from the backend. During verification like we upload tables, views, etc. onto cloud, we can download them in a similar way using a utility called ExecSqlizer. Also, multiple ETLs have been implemented that can help in populating tables directly onto the cloud using configurable parameters and environment variables. The job can be triggered from the SSISDB catalogs that triggers the Master package in the SSIS solution. The time taken for executing queries on database for millions of records takes almost 2hours for approximately 60GiBs of data, whereas after using our implemented technology, we can achieve the same in less than an hour with reduce file size of approx. 15-20 GiBs and also perform different querying and analysis based on the data populated. More on data processing and parallelism is explained in the research hereafter.

XIII. PARALLEL PROCESSING AND BETTER PERFORMANCE WITH DATA ON CLOUD

It is not unreasonable to say that modern datacentres and modern supercomputers are like twins separated at birth. Both are massively parallel in design, and both are organized as a network of communicating computational nodes. They both execute applications that are designed to exploit massive amounts of parallelism. Their differences lie in their evolution. Massively parallel supercomputers have been designed to support computation with occasional bursts of input/output and to complete a single massive calculation as fast as possible, one job at a time. In contrast, data centres direct their power outward to the world and consume vast quantities of input data.

Parallelism can be exploited in cloud computing in two ways. The first is for human access. Cloud applications are designed to be accessed as Web services, so they are organized as two or more layers of processes. One layer provides the service interface to the user's browser or client application. This "Web role" layer accepts users' requests and manages the tasks assigned to the second layer. The second layer of processes, sometimes known as the "worker role"

layer, executes the analytical tasks required to satisfy user requests.

The second way in which parallelism is exploited involves the nature of the data analysis tasks undertaken by the application. In many large data analysis scenarios, it is not practical to use a single processor or task to scan a massive dataset or data stream to look for a pattern—the overhead and delay are too great. In these cases, one can partition the data across large numbers of processors, each of which can analyze a subset of the data. The results of each "sub-scan" are then combined and returned to the user.

This "map-reduce" pattern is frequently used in datacentre applications and is one in a broad family of parallel data analysis queries used in cloud computing. Web search is the canonical example of this two-phase model. It involves constructing a searchable keyword index of the Web's contents, which entails creating a copy of the Web and sorting the contents via a sequence of map-reduce steps. Three key technologies support this model of parallelism: Google has an internal version, Yahoo! has an open source version known as Hadoop, and Microsoft has a MapReduce tool known as DryadLINQ.

Parallel programming approach addresses the parallel processing of data on data-intensive systems. Programming abstractions including models, languages and algorithms allow a natural expression of parallel processing. Design of data intensive computing platforms provide high levels of readability, efficiency, availability and scalability. Identifying applications can exploit this computing paradigm and determine how it should evolve to support emerging data intensive applications.

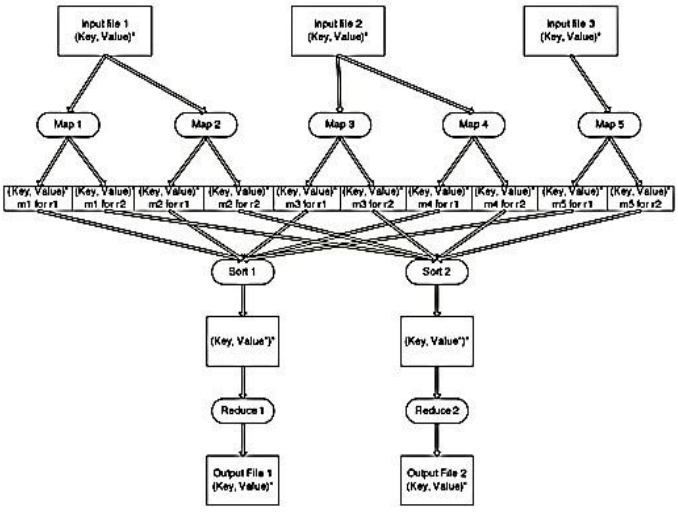

Fig. 4 MapReduce Key-Value Processing

With all the above promising features of cloud, we can assume that the accessibility to computation power is no longer a barrier for the users who need to perform large scale data/compute intensive applications. However, to perform such computations, two major pre-conditions need to be satisfied:

(i) the application should be parallelizable to utilize the available resources; and

(ii) there should be an appropriate parallel runtime support to implement it.

Feature	Hadoop	Dryad & DryadLINQ	CGL-MapReduce
Programming Model	MapReduce	DAG based execution flows	MapReduce with *Combine* phase
Data Handling	HDFS	Shared directories/ Local disks	Shared file system / Local disks
Intermediate Data Communication	HDFS/ Point-to-point via HTTP	Files/TCP pipes/ Shared memory FIFO	Content Distribution Network (NaradaBrokering[23])
Scheduling	Data locality/ Rack aware	Data locality/ Network topology based run time graph optimizations	Data locality
Failure Handling	Persistence via HDFS Re-execution of map and reduce tasks	Re-execution of vertices	Currently not implemented (Re-executing map tasks, redundant reduce tasks)
Monitoring	Monitoring support of HDFS, Monitoring MapReduce computations	Monitoring support for execution graphs	Programming interface to monitor the progress of jobs
Language Support	Implemented using Java Other languages are supported via Hadoop Streaming	Programmable via C# DryadLINQ provides LINQ programming API for Dryad	Implemented using Java Other languages are supported via Java wrappers

Fig. 5. Comparison of features supported by different cloud technologies.

From a programming perspective model, the MapReduce abstract model is based on the following concepts: Iteration over the input, computation of key/value pairs from each piece of input, grouping of all intermediate values by key, iteration over resulting groups, and reduction of each group.

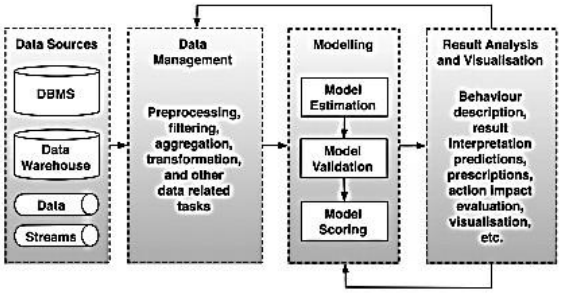

Fig. 6 An overview of analytics workflow of Big Data

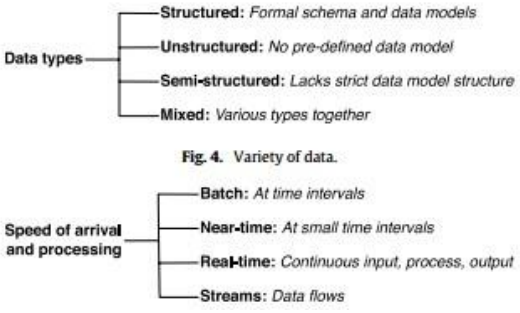

Fig. 4. Variety of data.

Fig. 7 Processing speeds of data

Considering data velocity, it is noticed that, to complicate matters further, data can arrive and require processing at

different speeds, as illustrated in Fig. 7. Whist for some applications, the arrival and processing of data can be performed in batch, other analytics applications require continuous and real-time analyses, sometimes requiring immediate action upon processing of incoming data streams.

Get It Right, Go Live, And Grow Flexibly!

XIV. REASONS TO PREFER OUR SYSTEM WITH RESPECT TO OTHERS

Some of the major issues faced by companies promoting data on cloud are related to: data security and compliance, data recovery, noisy neighbour, data management, adequate understanding of the cloud-based service provider and other such vulnerabilities. As most of the companies are moving or moved their data on cloud, being a future expectation; it is more likely that everyone must know which system to prefer and how to get their best from it.

While researching, we came across a paper 'Google for Work Security and Compliance Whitepaper', which stated Google fully understands the security implications and hence it delivers services to deliver better security than traditional on-premises solutions.

While we talk about Oracle 12c with Data on Cloud, there are multiple benefits offered for free and a payment basis. Although, problems such as data security, redundancy and data recovery are handled well, issues like noisy neighbour still persist. The noisy neighbour effect causes other virtual machines and applications that share the infrastructure to suffer from uneven cloud network performance.

The ways our system can avoid this is: to use a *bare-metal cloud*. The bare-metal cloud runs one application at a time directly on the hardware, which creates a single-tenant environment and eliminates noisy neighbours. While single tenant environments avoid the noisy neighbour effect, they do not solve the problem. Infrastructure over-commitment, or when an environment is shared by too many applications, limits overall cloud performance.

The most common issues that companies are facing today are:

i. lack of understanding that they are already in the cloud and they should have already been protecting themselves accordingly,
ii. lack of trust in cloud security,
iii. to protect their environment from data breaches or leaks and/or malicious attacks,
iv. lack of analysis on actual application code for vulnerabilities and data leaks.

Cloud services are like real clouds: They have holes; they're not solid. This is relatively a new technology and it needs time to arrive at a point where all possible risks can be laid out and analysed — for individual vendors as well as the cloud service industry as a whole.

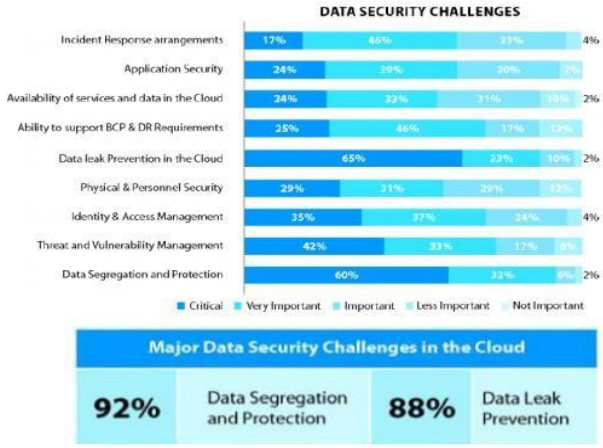

Fig. 8 Data Security Challenges

By encryption, our system offers a better solution to secure information on the cloud. Before uploading data into the cloud the users are suggested to verify whether the data is stored on backup drives and the keywords in files remain unchanged. Use RSA or 3-DES algorithms for higher security and MD5 hash calculations. A data driven framework has been designed for secure data processing and sharing between cloud users. Network based intrusion prevention system is used to detect threats in real-time. It is not important that which system is best and that you should go along with, but instead realize the importance of the needs of each system and accordingly choose the one you require.

Also, our system makes use of only 60-80% of the CPU utilization. Tokenization plays and important role in accelerating the performance and improves the overall processing speed on the cloud. The overall approach initiates with log processing, analysis and log management and goes on to mining data. Classification, Clustering, Curating, Association and decision making are some of the algorithms and techniques that are performed for knowledge discovery from databases. Data Curating is performed to organize and integrate data collected from various sources, annotate, publish and present data. For data recovery, we create 3 backups of data and use the RAID concept.

Tools that are used include PowerShell scripts that help in uploading and downloading data from cloud, Visual Studio with GIT integration for multiple people to work on one set of code, SQL Server Data Tools for creating and executing SSIS packages, SQL Server Profiler and SQL Analyser to analyze the query performance and reduce overall costs. Recently, we performed a small POC for integrating Cortana with Power BI, which can help us to communicate with the AI and indirectly get trends of data. We can also question it to get 10+ year old data and it can intelligently retrieve the data in form of reports visualization.

The overall architecture of our proposed system is depicted as follows –

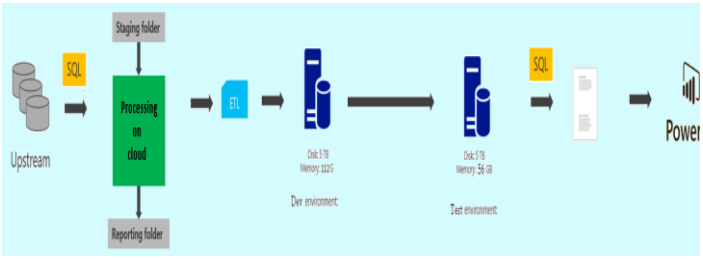

Fig. 9 Proposed system architecture

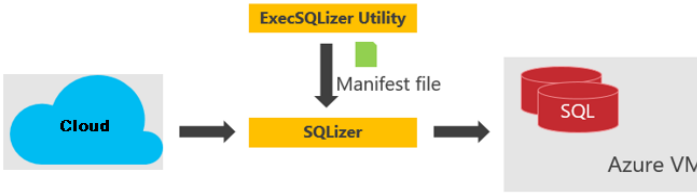

Fig. 10 Downloading data onto Azure VM

XV. CONCLUSIONS AND FUTURE SCOPE

The amount of data currently generated by the various activities of the society has never been so big, and is being generated in an ever increasing speed. This Big Data trend is being seen by industries as a way of obtaining advantage over their competitors: if one business is able to make sense of the information contained in the data reasonably quicker, it will be able to get more customers, increase the revenue per customer, optimise its operation, and reduce its costs. Nevertheless, Big Data analytics is still a challenging and time demanding task that requires expensive software, large computational infrastructure, and effort. Cloud computing helps in alleviating these problems by providing resources on-demand with costs proportional to the actual usage. Furthermore, it enables infrastructures to be scaled up and down rapidly, adapting the system to the actual demand.

There are plenty of solutions for Big Data related to Cloud computing. Such a large number of solutions have been created because of the wide range of analytics requirements, but they may sometimes, overwhelm non-experienced users. Analytics can be descriptive, predictive and prescriptive. Big Data can have various levels of variety, velocity, volume and veracity. Therefore, it is important to understand the requirements in order to choose appropriate Big Data tools. It is also clear that analytics is a complex process that demands people with expertise in cleaning up data, understanding and selecting proper methods and analysing results. Tools are fundamental to help people perform these tasks. In addition, depending on the complexity and costs involved in carrying out these tasks, providers who offer Analytics as a Service or Big Data as a Service, can be a promising alternative compared to performing these tasks in-house. Cloud computing plays a key role for Big Data, not only because it provides infrastructure and tools, but also because it is a business model that Big Data analytics can follow (e.g. Analytics as a Service (AaaS) or Big Data as a Service (BDaaS)).

However, AaaS/BDaaS brings several challenges because the customer and provider's staff are much more involved in the loop than in traditional Cloud providers offering infrastructure/platform/software as a service.

We expect that the applications developed using cloud technologies will work fine with cloud resources, because the Milliseconds-to-seconds latencies that they already have under the MapReduce model will not be affected by the additional overheads introduced by the virtualization.

As we enter the period in which all of science is being driven by a data explosion, cloud computing and its inherent ability to exploit parallelism at many levels has become a fundamental new enabling technology to advance human knowledge.

In this research paper, we proposed the theoretical and practical aspects of why to move over all on-premise database and related data to cloud. After understanding the hot and cold areas for having Big Data on cloud, we certainly believe that sooner most organizations will be moving them over.

As we are already working on these aspects, and learning simultaneously, we recommend that before you begin with the actual implementation and transfer process, it would be ideal to go through some of the webinars and seminar/conferences conducted by the cloud developers. This will give you an in-depth knowledge of the performance, parallel-processing options, storage, security, data recovery and backups, fault-tolerance and other regions to be analysed while processing data on cloud.

Don't wait, start right now – Future of Cloud is not far away!

REFERENCES

[1] *http://research.microsoft.com/en-us/collaboration/fourthparadigm/4th_paradigm_book_part3_gannon_reed.pdf*
[2] *http://cgl.soic.indiana.edu/publications/cloudcomp_camera_ready.pdf*
[3] *http://link.springer.com/chapter/10.1007%2F978-3-642-12636-9_2*
[4] "Handbook on cloud computing" by Borko Furht, Armando Escalante
[5] Moving to the Cloud: Developing Apps in the New World of Cloud Computing By Dinkar Sitaram, Geetha Manjunath *http://www.cloudbus.org/papers/BDC-Trends-JPDC.pdf*
[6] *http://www.computerweekly.com/feature/how-to-move-data-andapplications-in-the-cloud*
[7] *http://searchenterpriselinux.techtarget.com/tip/installing-and-setting-upopenerp-60*
[8] *http://searchenterpriselinux.techtarget.com/video/advantages-of-a-hybridcloud-infrastructure*
[9] *http://searchenterpriselinux.techtarget.com/definition/software-as-aservice*
[10] *http://searchcloudstorage.techtarget.com/feature/cloud-integratedstorage-appliances-link-on-premises-storage-to-cloud*
[11] *http://searchdatamanagement.techtarget.com/definition/extracttransform-load*
[12] *http://searchcloudstorage.techtarget.com/feature/cloud-storage-five-bestpractices-for-moving-to-the-cloud*
[13]*http://searchcrm.techtarget.com/news/2240017173/sugarcrm-software-demonstration*
[14]*http://searchbusinessintelligence.techtarget.in/tip/etl-tool-buying-guide*
[15]*http://searchsalesforce.techtarget.com/definition/Salesforcecom*
[16]*http://whatis.techtarget.com/definition/OpenStack*
[17]*http://whatis.techtarget.com/definition/multi-tenancy*