ISSN(Online) : 2319 -8753
ISSN (Print) : 2347 -6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 10, October 2015

An Appraisal to Overhaul Big Data Processing in Cloud Computing Environments

Vamsi Krishna Myalapalli[1], Karthik Dussa[2], Thirumala Padmakumar Totakura[3]

Open Text Corporation, Mind Space IT Park, Hitec City, Hyderabad, India

ABSTRACT: In the contemporary world it is inevitable to minimize accumulation of exponential data. Apart from actual data, data formats are also growing. Explosive data growth by itself, does not accuratelydescribe how data is changing; the format and structureof data are changing. Rather than being neatly formatted,cleaned, and normalized data in a corporate database, thedata is coming in as raw, unstructured text.The exploration of this paper could serve as a benchmarkingand management tool for overhauling Big Data management practices andprocessing. Experimental fallouts of our investigation advocate that maintainability is enriched.

KEYWORDS: Big Data Optimization; Efficient Big Data Management in Cloud; Big Data Analysis; Big Data Analytics; Big Data Best Practices.

I. INTRODUCTION

Typically Big Data is simply a large amount of data, but it is defined by more than just size.Big Data are high-volume, high-velocity, and/or high-variety information assets that require new forms of processing to enable enhanced decision making, insight discovery and process optimization.

Big Data is changing how businesses do business. Data is growing at an explosive rate, coming into the company from different areas, and in myriad formats. Social media, sensor data, spatial coordinates, and external data resource providers are just some of the new data vectors companies must now address. The result is that existing analytic and Business Intelligence (BI) practices must be rethought in the context of Big Data.

Powerful analytic platforms allow data analysts to rapidly build and deploy analytic applications to business decision makers. Big Data is changing how we manage data and how we use it in our businesses. Big Data comes in many forms, and from new sources such as mobile devices (e.g. smart phones), scientific sensors, and the cloud, and it's coming at fire hose speed. Smart companies realize that the rules of data are changing, and they need to improve how they manage Big Data to remain relevant and competitive in the marketplace.

II. BACKGROUND AND RELATED WORK

There exists lot of data, which comes into the system rapidly,and it comes from many different sources in many differentformats.Much of the data generated by new technology is unstructuredor in a semi-structured data format that makes it moredifficult to manage and process.

Optimizing Big Data Processing by Integrating Front End [1] proposed a model which integrates a versatile Front End Module to Database Systems.All the data that passes to and fro the database must go through this module. User specific security modules can also be loaded into this module.

This paper proposes efficient big data management practices and processing methodologies.

III.PROPOSED MODEL

Leveraging any technology takes some effort, but the payoffs are huge. Knowing a few tips, tricks, and guidelines to make our efforts easier and maximize the benefit is always advisable. This paper looks at ways to ensure that our Big Data experience is successful and that we gain the largest competitive advantage possible out of our Big Data Analytics investment.

ISSN(Online) : 2319 -8753
ISSN (Print) : 2347 -6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 10, October 2015

1) ***Ensure That the Big Data Analytics Platform Combines Cloud Experience with Sophisticated Analytics:***
Cloud computing provides the consumer with multiple benefits: rapid deployment, easy access from anywhere, security, and reduced IT overhead. However, it isn't sufficient to only access lightweight, low-capability BI and dashboards located in the cloud. Business decision makers require high-performance analytics, but with the convenience and cost benefits of the cloud. Make sure that the sophisticated, analytic platforms are cloud friendly. Don't assume that because the analytics are powerful, they are not suitable for the cloud. On the contrary, the most powerful processing tools are the best candidates for cloud computing. To leverage the limitless processing power of the cloud to run our sophisticated analytic tools so they always will be easily accessible to the Data Artisans and business decision makers.

2) ***Access All the Relevant Data to Make the Best Possible Strategic Business Decisions:***
Data comes in many different forms. At one end of the spectrum, data can be structured neatly inside a well-defined corporate database. At the opposite end, the data can be social media with no structured format at all. The data also can be somewhere in between as semi-structured data. Additionally, the data can come from traditional IT corporate data warehouses, desktop workstation documents and spread-sheets, automated smart devices and sensor equipment, or the cloud. Regardless of source or format, all relevant data has value to the decision maker and must be accessible to be useful. A decision based on bad, incomplete, or unrepresentative data is likely to be a bad decision. Use analytic tools that access all the relevant data available regardless of structure, source, or format. The analytic platform must blend the data seamlessly from the various sources and structures. Furthermore, ensure that the software tools implemented will access the myriad of data sources quickly and easily so that critical business decisions are not delayed.

3) ***Use a Single Platform for the Complete Analytic Process:***
The journey from accessing a multitude of different large data sources to quickly answering key business questions is indeed technically complex. Historically, highly specialized hardware and software to support data management, Extract Transform Load (ETL), integration, access, analysis, reporting, and presentation were common. Each component was highly specialized and required a specific skill set and integration process to transform raw data to valued business decisions. Besides cost and complexity, this process simply takes too long in today's business environments where opportunities exist for only short periods of time before they are seized by competitors. Be sure to select a platform that wraps up the complete end to-end analytic process, not just a mix of complex, disjointed components. The platform selected must recognize the workflow associated with analytics and manage that workflow and associated processes from a unified, easy-to-understand user interface.

4) ***Leverage an Analytic Platform to Access and Make Business Sense of Big Data:***
Gleaning real value out of Big Data is critical, otherwise data collection and processing is pointless. However, getting real value quickly is not a trivial task. The complexities associated with analytics have historically limited access to value of Big Data. To get real value from Big Data, analytic platform needs to do following:
a) Ensure availability of all types and formats of Big Data but without long wait times for access, integration, and processing. Big Data sources must be easily and seamlessly integrated with other, more-traditional data sources without requiring highly skilled technical staff.
b) Leverage the existing Data Artisans, without a massive retraining effort. The target platform needs to be simple and intuitive enough for Data Artisans to use quickly and effectively.

5) ***Move from Social Media Feedback to Real Business Knowledge:***
Many companies monitor social media sources yelp for consumer comments. Seeing a number of "Likes" for a product or service is a powerful indicator, but what does it actually mean in terms of real business value. The solution is integration of social media data with more traditional and structured data sources to gather a complete picture of the consumer environment. However, integrating the geospatial characteristics of the social data with more concrete data from point-of-sale and customer loyalty programs will quantify the true value of those social media inputs. Be sure that the analytic toolset can support this kind of integration so that we capture a complete, quantitative view of the data and financial value.

ISSN(Online) : 2319 -8753
ISSN (Print) : 2347 -6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 10, October 2015

6) Gain Value from Big Data Sooner, Rather than Later:

Better data yielding better decisions is an easy concept for people to understand. Translating those better decisions into competitive advantage and increased sales is also an easy sell for most business leaders. The problem occurs when the development and implementation path proves to be long, tedious, and expensive. Frequently, a complex analytic solution deployed in-house by an already overworked IT department can take many months or even years to yield quantifiable benefits for everyday business users.

Efficient Platforms avoid this pitfall by being immediately usable by the Data Artisans and decision makers. Leveraging access to external Big Data sources, integration with internal data stores, and an easy-to-use single toolset bring the value of Big Data to the company immediately. Rather than embarking on a long IT project, decision makers and Data Artisans are actively using Big Data Analytics with quantifiable results in a very short timeframe.

7) Recognizing Where the Business and Customers are Located:

Operating without a profile of the customers is obviously a path to failure. Demographic data has proven itself critical, and many demographic data sources have been available for years. However, one of the most important data attributes historically has been overlooked due to technical challenges; fortunately, the use of geospatial data is now possible by everyone.

Geospatial intelligence tells us where our customers are, where they visit, where they buy our products, and where competitors are located. Social media tags data with geospatial tags and analytic processing tools take advantage of these tags to identify and analyse this data. Ensure that spatial intelligence is a part of any analytic processing toolset we utilize so that this data will enhance critical decision making.

8) Valuing our Data Artisans and What They Bring to the Table:

The most important component in any system is always the people, and the same is true especially in BI and analytic processing. Of the staff, the most critical folks are the Data Artisans. They are the business experts, typically attached to a business unit, who truly understand our business and what data is critical. They understand IT, but technical knowledge is not how they bring value. The ability to know what data the decision maker needs, where to find that data, and how it needs to be analysed and processed is critical, game-changing skill set of the Data Artisans.

As with any advantage, a smart business leader wants to maximize that advantage for maximum impact. The way to make the most of our Data Artisans is to empower them and to provide the tools that best support their work. Agile, powerful, and easy-to-use platformstrategic analytics are the best way to leverage the inherent knowledge of our Data Artisans.

9) Shift in Processing Due to Big Data:

Traditionally, large datasets would reside on a corporate mainframe or in a data warehouse in a well-defined format, often managed by an advanced Relational Database Management System (RDBMS). This is a tried-and-true configuration, but it does not reflect the changing nature of Big Data. As the data has changed, so must how it is processed.

Traditional (Business Intelligence) BI tools that rely exclusively on well-defined data warehouses are no longer sufficient. A well-established RDBMS does not effectively manage large datasets containing unstructured and semi-structured formats. To support Big Data, modern analytic processing tools must

a) Shift away from traditional, rearward-looking BI tools and platforms to more forward-thinking analytic platforms.
b) Support a data environment that is less focused on integrating with only traditional, corporate data warehouses and more focused on easy integration with external sources.
c) Support a mix of structured, semi-structured, and unstructured data without complex, time-consuming IT engineering efforts.
d) Process data quickly and efficiently to return answers before the business opportunity is lost.
e) Present the business user with an interface that doesn't require extensive IT knowledge to operate.

Fortunately, IT vendors and the IT open source community are stepping up to the challenge of Big Data and have created tools that meet these requirements. Popular software tools include

ISSN(Online) : 2319 -8753
ISSN (Print) : 2347 -6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 10, October 2015

a) **Hadoop:** Open-source software from Apache Software Foundation to store and process large non-relationaldata sets via a large, scalable distributed model.
b) **NoSQL:** A class of database systems that are optimized to process large unstructured and semi-structured data sets.

10) Changing Focus with Big Data:

Unlocking the value in data is the key to providing value to the business. Too often IT infrastructure folks focus on data capacity or throughput speed. Business Intelligence vendors extol the benefits of executive-only dashboards and visually stunning graphical reports. While both perspectives have some merit, they only play a limited role in the overall mission of bringing real value to those in the company who need it.

Value is added by using an approach and platform to bring Big Data into the hands of those who need it in a fast, agile manner to answer the right business questions at the right time. Knowing what data is needed to answer questions and where to find it is critical; having the analytic tools to capitalize on that knowledge is even more critical. It is through those platforms that real value is realized from Big Data.

11) Implementing Big Data Analytics within an Organization:

Technology alone doesn't generate real value from Big Data. Data analysts, empowered with the right analytic technology platform, humanize Big Data, which is how companies realize value. Analytic platforms make extracting value from Big Data possible. Important benefits to businesses that the Strategic Analytics platform provides include.

a) Improving the self-sufficiency of decision makers to run and share analytic applications with other data users. Data analysts who understand the business should develop good analytic applications that are shared for everyone's benefit.
b) Injecting Big Data into strategic decisions without waiting months for an IT infrastructure and data project. Getting the data into the hands of decision makers so that businesses can identify and capitalize on opportunities.
c) Delivering the power of predictive analytics to everyone, not just a few executive decision makers far removed from operations. Ensuring that the right data is readily available to all authorized parties leads to making the best possible decisions.

12) Blending Data from Multiple Sources:

The nature of Big Data is large data, usually from multiple sources. Some data will come from internal sources, but increasing data is coming from outside sources. These outside sources include

a) Social media data feeds.
b) Point of Sale and customer loyalty tracking programs.
c) Government agency sources such as census data.
d) Spatial data from mobile devices and satellite mapping feeds.
e) Consumer demographic data brokers, such as Experian.
f) Any number of public, private, or community clouds.

Data blending is the process of combining multiple heterogeneous data sources and blending them into a single, usable analytic dataset. The purpose of data blending is to create analytic datasets to answer business questions using data that is not bound by the control and lengthy timelines of traditional IT processes. An example of data blending is when the data analyst integrates packaged, external data from the cloud with internal data sources to create a very business-specific analytic dataset.

Data blending is important, and any Big Data Analytics platform must support this function. With data blending, the complete scope of Big Data becomes available to the data analyst.

13) Focussing on Context, Not Just Integration:

Data integration is a big piece of the picture that often receives a great deal of attention, but it is not the most important aspect of working with Big Data. Putting Big Data into context with all our other sources of data is what is important. Company's internal databases, corporate data warehouses, and myriad of spread-sheets and

ISSN(Online) : 2319 -8753
ISSN (Print) : 2347 -6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 10, October 2015

documents hold a wealth of information. The limitless sources of Big Data need to be pared down to what is relevant and be placed in context with the internal data sources.

Using analytic tools allows us to place our Big Data into context with internal data sources promoting maximum efficiency and agility.

Data Integration best practices:

a) Integrate Big Data with all required data including local data, corporate data stores and data from cloud applications.
b) Integrate packaged market and customer data.
c) Work to process the data without explicit, lengthy IT integration involvement whenever possible.
d) Allow data analysts to build analytic applications that incorporate complete context for decision making.

Focusing on context first with integration second allows for a better, more confident decision-making experience without the risk of becoming bogged down in a long integration project.

Use all the relevant data in an agile manner. We need to understand what data we really need and then leverage it quickly.

14) Combining Big Data with Spatial Data:

Spatial data is defined as data that identifies the geographic locations of objects or features on Earth. The widespread deployment of Global Positioning Satellite (GPS) devices has coincided with the practice of geo-tagging which has been growing at an explosive rate. Geo-tagging is the process of adding spatial metadata to generated content such as photographs, SMS texts, Rich Site Summary (RSS) feeds, and social media posts. This process provides a continual stream of data identifying where people are and what they are doing and is a major part of the growing wave of Big Data. Analytic platforms that understand and process this type of data without additional, specialized technologies give the business user a powerful edge.

The Strategic Analytics platform with native spatial capability leverages spatial attributes in Big Data to:

a) Deliver deep spatial and location understanding as part of the unified workflow to answer contemporary business questions more effectively.
b) Provide a more complete understanding of transactions and market and customer interactions.
c) Improve the outcome of critical customer, market, and investment decisions without the requirement for complex and specialized GIS software.

Taking spatial data attributes into context with our otherdata sources is the best method to leverage our Big Datausing the analytic platform. Blending the spatial dataprovided by geo-tagging with the internal, customer, anddemographic data resources opens the doorway to realizingthe full value of Big Data Analytics in decision making.

15) Leveraging External Data Provider Resources:

Data aggregation service providers are companies that assist clients by capturing and leveraging data about their target market, customer base, and industry. These companies use that data to gain a competitive advantage, for a price.

External, packaged market data provides a company with customer and market context that individual companies cannot obtain. These data sources provide companies with information that cannot easily be recreated in a format tailored to their requirements.

Powerful analytic platforms need to integrate easily with external data providers. Fast and simple access to valuable third-party data sources allows data analysts to blend internal corporate data with external dataresources for the best possible analytic results and decisionmaking.

16) Gaining Real Business Value from Predictive Analysis:

Smart businesses apply predictive analytics to their decisionmakingprocess to obtain large benefits and advantages intheir marketplace. Examples include

a) Taking a proactive rather than a reactive position basedon real-time data trends and predictions to ensure theleadership position in the market.

ISSN(Online) : 2319 -8753
ISSN (Print) : 2347 -6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 10, October 2015

b) Identifying and responding to trends on social media outletsto take advantage of opportunities. For example, if acompany detects rapidly increasing interest in a product,the company proactively can increase inventory in theimpacted geographic area as identified by the spatial data.
c) Quantifying and leveraging the actual value of socialmedia comments on the business. Blend the context ofBig Data with user sentiment analysis and quantitativedata to gain an accurate picture of the environment andmarket direction.

Predictive analysis is a powerful tool for decision makers.

17) Considering Consumerization of Big Data Analytics:

Consumerization is widely defined as the practice of taking an IT product that is popular in the consumer market space and introducing it into the workplace. Consumerization of Big Datacan be achieved by:

a) Leveraging cloud technology as a social experience for decision makers to run and share their applications anytime and from anywhere. This contributes to the self-sufficiency and user independence.
b) Bringing Big Data context into decision-making process immediately without waiting for a large Project.
c) Placing power of predictive analytics into the hands of every decision maker without adding complexity.

Consumerization of Big Data allows for better decisions to be made faster and with less complexity by those who need the data the most.

18) Publishing Data and Analytics to Cloud Service:

After the data analyst has created an analytic application, it must be deployed as quickly and easily as possible for use by the business users. The optimal deployment environment is the cloud, from which applications:

a) Are available to designated users regardless of location.
b) May be shared with other data analysts or end users in the community.
c) Are secure so only authorized users can access the application and resulting data.
d) Are fast, highly available, and independent of traditional IT infrastructure.

19) Focusing on ConsumingApplications:

Big Data Analytics consumerization is important because itmakes analytics accessible and easy for the end user. Severalkey considerations exist for successful consumerization ofapplications:

a) Easy access to Big Data for end users is essential. Accessmust be easily obtained without complex IT involvement.
b) Consumerization should focus on the decision-makerexperience and consuming applications. Consider theperspective of the decision maker when developingapplications.
c) Specific devices such as mobile devices and smart phones should not be the primary focus. Technology is continually changing and the underlying technological stack should be abstracted from the users' view.
d) All relevant data sources should be consumed. This is where data analysts show their value by excluding extraneous data while not missing what is critical.
e) Easy-to-use parameterized wizards and interfaces that solicit user input to enhance and socialize the user experience.This is how user "drives" application, so interface must make sense from a user's perspective.
f) Complex and sophisticated algorithms should be contained within predictive analytic tools without forcing the user to have a deep understanding of the behind-the-scenes technical details. Hiding complexity is achieved by prebuilt analytic tools, and formulas at the data analyst's fingertips.

Shifting the focus from IT and complex data design to intelligently consuming applications so Big Data is easily accessible to business users and decision makers will ensure a successful Big Data Analytics implementation.

20) Best Platform for Strategic Analytics:

Strategic analytics is the sophisticated analysis used to make critical decisions that drive business strategy and growth. This analysis is a direct contrast to the static, prebuilt, and predefined reports and executive dashboards focused on past data and performance. Modern strategic analytics products look at the current environment and use Big Data context with advanced analytic tools to gain insights into the future. This is important because the forward-looking approach offers the greatest avenue to identify and capitalize on key business opportunities.

ISSN(Online) : 2319 -8753
ISSN (Print) : 2347 -6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 10, October 2015

21) Structure Big Data Environments around Analytics, not Ad hoc Querying or Standard Reporting:

Every step in the data pathway from original source to analyst's screen must support complex analytic routines implemented as user defined functions (UDFs) or via a metadata driven development environment that can be programmed for each type of analysis. This includes loaders, cleansers, integrators, user interfaces, and finally Business Intelligence tools. This practice does not recommend repudiating our existing environment, but rather extending it to support the new demands of analytics.

22) Embrace Sandbox Silos and Build a Practice of Productionizing Sandbox Results:

Allow data scientists to construct their data experiments and prototypes using their preferred languages. Then, after proof of concept, systematically reprogram and/or reconfigure these implementations.

23) Architectural Implementation for Big Data:

Plan for a logical "data highway" with multiple caches of increasing latency. Physicallyimplement only those caches appropriate for the environment. The data highway canhave as many as five caches of increasing data latency, each with its distinct analytic advantages and trade-offs.

Figure 3.1: Architectural Implementation for Big Data

Raw Source Applications: Credit card fraud detection, immediate complex eventprocessing (CEP) including network stability and cyber-attack detection.

Real time Applications: Web page ad selection, personalized price promotions,online games monitoring, various forms of predictive and proactive monitoring.

Business Activity Applications: Low latency KPI dashboards pushed to users,trouble ticket tracking, process completion tracking, "fused" CEP reporting,customer service portals and dashboards, and mobile sales apps.

Top Line Applications: Tactical reporting, promotion tracking, mid-coursecorrections based on social media buzz. "Top line" refers to the common practiceby senior managers of seeing a quick top line review of what has happened inthe enterprise over the past 24 hours.

EDW and Long-Time Series Applications: All forms of reporting, ad hoc querying,historical analysis, master data management, large scale temporal dynamics,and Markov chain analysis.

Each cache that exists in a given environment is physical and distinct from the othercaches. Data moves from the raw source down this highway through ETL processes.There may be multiple paths from the raw source to intermediate caches. Forinstance, data could go to the real time cache to drive a zero latency-style userinterface, but at the same time be extracted directly into a daily top line cache thatwould look like a classic Operational Data Store (ODS). Then the data from this ODScould feed the EDW. Data also flows in the reverse direction along the highway.

24) Use Big Data Integration to Build Comprehensive Ecosystems that Integrate Conventional Structured RDBMS Data, Paper based Documents, Emails, and in-house Business-Oriented Social Networking:

One of the potent messages from big data is theability to integrate disparate data sources of very different modalities.Capture all this information, dimensionalize it, use it in the course of business, and thenback it up and archive it.

ISSN(Online) : 2319 -8753
ISSN (Print) : 2347 -6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 10, October 2015

25) Plan for Data Quality to be Better Further along the Data Highway:

This is the classic trade-off of latency versus quality. Analysts and business users must accept the reality that very low latency (i.e., immediate) data is unavoidably dirty because there are limits to how much cleansing and diagnosing can be done in very short time intervals.

26) Apply Filtering, Cleansing, Pruning, Conforming, Matching, Joining, and Diagnosing at the Earliest:

This is a corollary of the previous best practice. Each step on the data highway provides more time to add value to the data. Filtering, cleansing and pruning will reduce the amount of data transferred to the next cache and eliminates irrelevant or corrupted data. To be fair, there is a school of thought that applies cleansing logic only at analysis run time, because cleansing might delete"interesting outliers." Conforming takes the active step of placing highly administeredenterprise attributes into major entities like customer, product, and date. Theexistence of these conformed attributes allows high value joins to be made acrossseparate application domains. A shorter name for this step is "integration!"Diagnosing allows many interesting attributes to be added to data, including specialconfidence tags and textual identifiers representing behavior clusters identified by adata mining professional. Data discovery and profiling, assists in identifying data domains, relationships, metadata tags useful for search, sensitive data, and dataquality issues.

27) Implement Streaming Data Analytics in Selected Data Flows:

An interesting angle on low latency data is the desire to begin serious analysis on the data as it streams in, but possibly far before the data transfer process terminates. There is significant interest in streaming analysis systems which allow SQL-like queries to process the data as it flows into the system. In some use cases, when the results of a streaming query surpass a threshold, the analysis can be halted without running the job to the bitter end. An academic effort, known as continuous query language (CQL), has made impressive progress in defining the requirements for streaming data processing including clever semantics for dynamically moving time windows on the streaming data. An ideal implementation would allow streaming data analysis to take place while the data is being loaded at gigabytes per second.

28) Perform Big Data Prototyping on a Public Cloud and then Move to a Private Cloud:

The advantage of a public cloud is that it can be provisioned and scaled up instantly. The sensitivity of the data allows quick in-and-out prototyping, this can be very effective. Never leave a huge data set on-line with the public cloud provider over the weekend when the programmers have gone home. However, we should remember that in some cases where we are trying to exploit data locality with rack-aware MapReduce processes, we may not be able to use a public cloud service because they may not give us the data storage control we need.

29) Ensure 10x to 100x Performance Improvements, RecognizeParadigm Shift for Analysis at High Speeds:

The openness of the big data marketplace has encouraged hundreds of special purpose tightly-coded solutions for specific kinds of analysis. This is a giant blessing and a curse. Once free from being controlled by a big vendor's RDBMS optimizer and inner loop, smart developers can implement spot solutions that are truly 100 times as fast as standard techniques.

30) Separate Big Data Analytic Workload from Conventional EDW to Preserve EDW Service Level Agreements:

If big data is hosted in Hadoop, it probably doesn't compete for resources with the conventional RDBMS-based EDW. However, we should be cautious if big data analytics run on the EDW machine since big data requirements change rapidly and inevitably in the direction of requiring more compute resources.

31) Exploit Unique Capabilities of In-Database Analytics:

The major RDBMS players all invest significantly in in-database analytics. Once wepay the price of loading data into relational tables, SQL can be combined with analytic extensions in extremely powerful ways.

32) Dimensionalize the Data Before Applying Governance:

This is a challenge that big data introduces. We must apply data governance principles even when we don't know what to expect from the content of the data. We will receive data arriving at up to gigabytes per second,

ISSN(Online) : 2319 -8753
ISSN (Print) : 2347 -6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 10, October 2015

often as name–value pairs with unexpected content. Our best chance at classifying data in ways that are important to our data governance responsibilities is to dimensionalize it as fully as possible at the earliest stage in our data pipeline. Parse it, match it, and apply identity resolution on the fly.

Big Data Modelling Best Practices #(33-40):

33) ***Dimensionalization:***

Divide the data into dimensions and facts. Automated dimensioning is required to stay ahead of the high velocity streams of data. Incoming data should be fully dimensionalized at the earliest extraction step.

34) ***Integrate separate data sources with conformed dimensions:***

Conformed dimensions are the glue that holds together separate data sources, and allow them to be combined in a single analysis. Conformed dimensions are perhaps the most powerful best practice from the conventional EDW world that should be inherited by big data

35) ***Anchor all Dimensions with Durable Surrogate Keys:***

The first step in every data source is to augment the natural key coming from a source with an enterprisewide durable surrogate key. Durable mean that there is no business rule that can change the key. The durable key belongs to IT, not to the data source. Surrogate means that the keys themselves are simple integers either assigned in sequence or generated by a robust hashing algorithm that guarantees uniqueness. An isolated surrogate key has no applications content. It is just an identifier.

36) ***Expect to Integrate Structured and Unstructured Data:***

Big data considerably broadens the integration challenge. Much big data will never end up in a relational database; rather it will stay in grid. But once we are armed with conformed dimensions and durable surrogate keys, all the forms of data can be combined in single analysis. This step needs to take place at query time, not at data loadand structure time.The most flexible way to perform this integration is through datavirtualization, where the integrated data sets appear to be physical tables but areactually specifications similar to relational views where the separate data sources arejoined at query time. If data virtualization is not used, then the final BI layer mustaccomplish this integration step.

37) ***Track Time Variance with Slowly Changing Dimensions (SCDs):***

Tracking time variance of dimensions is an old and venerable best practice from the EDW world. Basically, it makes good on the pledge we take to track history accurately.

38) ***Get Used to Not Declaring Data Structures Until Analysis Time:***

One of the charms of bigdata is putting off declaring data structures at the time of loading into data grid. The data structures may not be understoodat load time. The data may have such variable content that a single data structureeither makes no sense or forces us to modify the data to fit into a structure. If wecan load data without declaring its structure, we can avoida resource intensive step.

39) ***Build Technology Around Name-Value Pair Data Sources:***

In many cases, we open the fire hose and discover unexpected or undocumented data content which we must nevertheless load at gigabytes per second. The escape from this problem is to load such data as simple name-value pairs.

40) ***Use Data Virtualization to Allow Rapid Prototyping and Schema Alterations:***

Data virtualization is a powerful technique for declaring different logical data structures on underlying physical data. Standard view definitions in SQL are a good example of data virtualization. Data virtualization can present a data source in any format the analyst needs. But data virtualization trades off the cost of computing at run time with the cost of ETL to build physical tables before run time.

ISSN(Online) : 2319 -8753
ISSN (Print) : 2347 -6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 10, October 2015

IV. CONCLUSION

As Big Data accumulates in many formats we should anticipate the deleterious consequences as well. With Big Data, the way to be agile (and therefore successful) is to spot trends, opportunities, and risks via analytic processing of data, and in modern times. Thus, if we want to be successful, wemust understand Big Data and how to quickly extract from it the business critical information that the business requires.

REFERENCES

[1] Vamsi Krishna Myalapalli andThirumala Padmakumar Totakura, "Optimizing Big Data Processing in Cloud by Integrating Front End to Database Systems", IEEE International Conference on Energy System and Application, October 2015, Pune, India.
[2] Vamsi Krishna Myalapalli and Srinivas Karri, "Optimizing Front End Applications and JavaScript", Pages 4010-4020, Volume 4 Issue 6, June 2015, International Journal of Innovative Research in Science, Engineering and Technology.
[3] Michael Wessler, "Big Data Analytics for Dummies", 1st Edition, John Wiley & Sons Publishing, 2013.
[4] Changqing Ji, Yu Li, Wenming Qiu, Uchechukwu Awada and Keqiu Li, "Big Data Processing in Cloud Computing Environments", IEEE International Symposium on Pervasive Systems, Algorithms and Networks, 2012.
[5] Vamsi Krishna Myalapalli andPradeep Raj Savarapu, "High Performance SQL", IEEE International Conference on Emerging Trends in Innovation and Technology, December 2014, Pune, India.

BIOGRAPHY

Vamsi Krishna completed his Bachelor's in Computer Science from JNTUK University College of Engineering, Vizianagaram. He is Microsoft Certified Professional (Specialist in HTML5 with JAVA Script & CSS3), Oracle Certified JAVA SE7 Programmer, DBA Professional 11g, Performance Tuning Expert, SQL Expert and PL/SQL Developer. He has fervor towards SQL & Database Tuning.He can be reached at vamsikrishna.vasu@gmail.com.

Karthik completed his Bachelor's in Computer Science from Vidya JyothiInstitute of Technology, Hyderabad. He is OpenText Certified Content Server Solution Developer, Content Server System Administrator, Oracle Certified JAVA SE7 Programmer and Microsoft Certified Windows Application Developer. He has fervor towardsApplication Development, OScript, JAVA and SQL programming.He can be reached at karthikdussa29@gmail.com.

Padmakumar completed his Masters in Computer Science from Andhra University, Visakhapatnam and PG Diploma in Business Administration from Symbiosis University, Pune. He is Brainbench Certified JSP Developer. He has worked on sundry technologies which include J2EE, J2ME, XML, SQL, Enterprise Service Bus Architecture, Model Driven Architecture, Service Component Development, Enterprise Application Integration and Web services. He has fervor towards Business Analytics, Parallel Processing, Database Development and Enterprise JAVA Development etc. He can be reached at thirumalakumar@gmail.com.