Global journal of multidisciplinary and applied sciences

Available online at www.gjmas.com

ISSN xxxx-xxxx ©2013 GJMAS

UTILIZATION OF BIG DATA IN OIL AND GAS INDUSTRIES USING HADOOP MAPREDUCE TECHNOLOGY AND HIVEQL

M.Sakthivadivel* ,N.Krishnaraj and P.Ramprakash

Assistant Professor Department of Information Technology Dr. Mahalingam college of engineering and technology Pollachi-642003, India

***Corresponding author*:** M.Sakthivadivel

ABSTRACT: Big Data is basically vast amount of data which cannot be effectively processed, captured, and analyzed by traditional database and search tools in reasonable amount of time. Big Data information explosion is mainly due to the vast amount of data generated by social media platform, data input from omni-channels, various mobile devices, user agents, multimedia data, and so on. Overall it is an expanding "Digital Universe". Big Data predominately revolve around 3V's: **V**olume, **V**elocity, **V**ariety. Big data plays a major role in oil and gas industries. The purpose of this research is to analyze the various problems faced by the Oil & Gas industries in monitoring the vast amount of data received from their various units and to provide a suitable solution for monitoring the data. Oil and Gas (O&G) companies – both the operator companies as well as oil field service providers, now have more upstream data than they are processed before. They are at the verge of managing the vast amount of data generated from exploration , drilling or production works. Management of this is essential for effective, productive, and on demand data insight is critical, for decision making within the organization. The major problem faced by industries is the maintenance of data in concern with Exploration and Production processes. In the oil and gas industry, organizations must apply new technologies and processes that will capture and transform raw data into actionable insight to improve asset value and yield while enhancing safety and protecting the environment. Well and field operations are instrumented to capture a holistic view of equipment performance and well productivity data including reservoir, well, facilities and export data. Leading, analytics-driven oil and gas organizations are connecting people with trusted information to predict business outcomes, and to make real-time decisions that help them outperform their competitors. But for many, these breakaway results remain out of reach. Despite the wealth of data and content available today, decision makers are often starved for true insight. IBM's big data platform provides a scalable, easy to use, secure information management and analytics platform for complex and large-scale analysis and economical storage of drilling production data. Successfully harnessing big data unleashes the potential to achieve three critical objectives:

1. Enhance exploration and production
2. Improve refining and manufacturing efficiency &
3. Optimize global operations

With the help of IBM frameworks the O&G companies can adopt Hadoop enabled Big Data solutions for creating integrated digital Oil Field strategy. Hadoop based solutions allow storing, processing, and analyzing these humongous logs on near real time basis. The crux of solution involves processing raw data in its native format to create aggregated views along with understanding of its relationships and patterns and thereby derive meaningful insight for quick decision-making related to reservoir and optimizing the data exploitation using Map Reduce paradigm. HiveQL is a scalable Data Warehouse solution available on Hadoop, which is similar to SQL syntax. Hive internally generates map-reduce jobs that can be executed on Hadoop clusters. Hive in-turn allows overcoming the learning curve associated with the Map-Reduce code generation. With the help of Hadoop and HiveQL framework the Oil & Gas industries can manage the Big data from various units of oil refineries thereby generating a platform

for efficient utilization of the required data related to the companies for creating the excellence of customers in the global market.

***Keywords*:** Volume, Velocity, variety, Hadoop, Hive, Oracle.

INTRODUCTION

Big Data is a large volume of data from various data sources such as social media, web, genomics, cameras, medical records, arial sensory technologies and information sensing mobile devices. Big Data includes structured, semi-structured and unstructured data. This unstructured data contains useful information which can be mined. Since 1980s per capital capacity to store information is increased into double the amount for every 40 months. In 2012, statistics say that 2.5 quintillion bytes of data are created per day. Moreover, digital streams that individuals create are growing rapidly. For example, most of the people are using camera on their own. Big Data are of high level volume, high velocity, and high variety of information that needs advanced method to process the Big Data. In addition, conventional software tools are not capable of handling Big Data. So Big Data requires extensive architecture. The following types of data re referred to as big data

1. Social data - Customer feedback forms for Customer Relationship Management in Social media sites such as Twitter, Facebook, Linkedln, etc.
2. Machine-generated data - Sensor readings, Satellite Communication.
3. Traditional enterprise data - Employee information, business product, purchase, sales, customer information and ledger information.

TRAITS OF BIG DATA

Big Data differs from other data in 5 dimensions such as volume, velocity, variety and value.

a. Volume: Machine generated data will be large volume of data.

b. Velocity: Social media websites generates large data but not massive. Rate at which data acquired from the social websites are increasing rapidly.

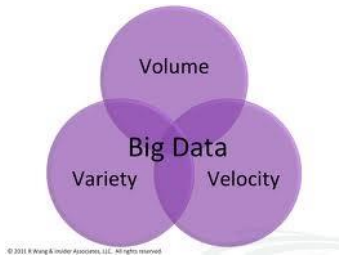
c. Variety: Different types of data will be generated when a new sensor and services.

d. Value: Even the unstructured data has some valuable information. So extracting such information from large volume of data is more considerable.

e. Complexity: Connection and correlation of data which describes more about relationship among the data.

f. Challenges: Storing and maintaining the Big Data is a challenging task. The following challenges needs to be faced by the enterprises or media when handling Big Data:

- Capture
- Duration
- Storate
- Search
- Sharing
- Analysis
- Visualization



WHY BIG DATA

Big Data is absolutely essential for the following intents:

- To spot business trends
- Determine quality of research
- To prevent diseases

- To link legal citation
- To combat crime
- To determine real time roadway communication system where the data is created in the order of exa bytes.

WHERE IT IS USED

Areas or fields where big data are created:

- Medicine, Metrology, Connectomics, Genomics, Complex Physics Simulation, Biological, Environmental research and Arial sensory system.
- Big Science, RFID, Sensor networks.
- Astormetry.net Project keeps eye on Astrometry group via flicker for new photos of the night sky. It analysis each image and identifies the celestial bodies.

BIG DATA ENABLED DIGITAL OIL FIELD

a. Oil And Gas Industry Overview

Oil and Gas companies – both operator companies as well as oil field service providers, now have more upstream data than ever before; to base their operational decisions relating to exploration, drilling or production. For this reason effective, productive and on-demand data insight is critical, for decision making within the organization.

However, a vision towards an integrated Exploration and Production data management platform, still remains a challenge as extraction of business-critical intelligence/insights from large volumes of data in a complex environment of legacy diverse systems and fragmented/decentralized solutions is a daunting task.

Some typical challenges of E&P data management are:

- Upstream focused applications are at a functional level. So, substantial time is spent in data collection about running reports for a given asset level i.e. for a single well or aggregated wells in a given location.
- A major number of applications are still non-PPDM based, which makes the reports and KPIs non-accurate at most times.\
- It is difficult to drive insights from unstructured data lying in multiple applications.
- It is difficult to run predictive analytics as data is spread out in multiple systems with lesser integrity and reference to master-level data.

NEED FOR DIGITAL OIL FIELD ENTERPRISE PLATFORM

An integrated digital oil field enterprise platform integrates E&P data from different project phases-seismic, drilling, well and production-into a single consolidated platform. Data indexing, storage, cleansing, clustering, migration, standardization, and analysis can be done from multiple data sources into an integrated platform and provide detailed insights at well level at any instant. This solution should leverage cloud infrastructure, an integrated workflow, an accelerated digitized solution framework based on MURA, hybrid data models, integration to multiple data sources, and a host of accelerators for data migration.

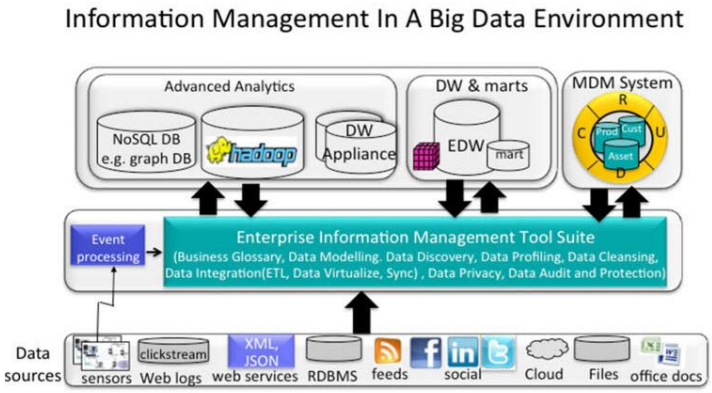
BIG DATA IN DIGITAL OIL FIEL

In oil and gas industry, traditional data warehousing solutions are facing challenges to capture, storage, and churn massive volume of datasets. The O&G companies can adopt Big Data solutions to maximize their business potential by driving a holistic view to gather valuable insights that can complement existing traditional BI offerings.

This consolidated Big Data enabled E&P data management platform should be designed to fit within an O&G operators or oil field service providers technology infrastructure and provides an on-demand and single view of a well at any instant and from anywhere. The platform should provide ready-to-use accelerators as well as interfaces with third-party Geologist and Geophysicists product suites as well as with customer data sources-be it structured, unstructured or real-time.

BIG DATA SOLUTION FOR DIGITAL OIL FIELD

O&G companies can adopt Hadoop enabled Big Data solutions for creating integrated Digital Oil Field strategy. Hadoop is a widely accepted open-source cost



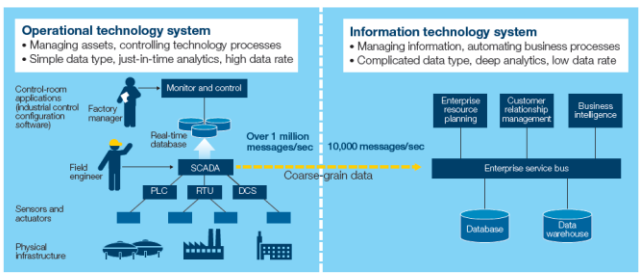
effective solution which provides map-reduce functionality for processing extremely large data sets on commodity severs. Hadoop based solutions allow storing, processing, and analyzing these humongous logs on near real time basis. The crux of the solution involves processing raw data in its native format to create aggregated views along with understanding of its relationships and patterns and thereby derive meaningful insights for quick decision-making related to reservoir and optimizing the data exploitation using Map Reduce paradigm. There is a widely acceptable adoption of Hive, which is a scalable data warehouse solution available on Hadoop, with HiveQL as query mechanism which is similar to SQL syntax. Hive internally generates map-reduce jobs that can be executed on Hadoop clustures. Hive in-turn allows overcoming the learning curve associated with the Map-Reduce code generation.

FUNCTIONAL OVERVIEW

As part of a faster transition strategy towards Digital Oil Field for integration, processing and analytics, Hadoop clusters can be leveraged along with data migration and business intelligence accelerators. An architecture view is depicted below strategizing a unified view of different Oil Wells managing semi/unstructured data of drilling and production phase, leveraging modeling/simulation techniques, and ready-to-deploy KPI configurations.

The high-level functional overview is stated below:

- Fetch the customer's E&P from various well in different phases – Each oil well generates around 10YB of data and in a reservoir there are multiple wells to be drilled and explored.
- This massive volume of multi structured logs is stored on a Hadoop infrastructure.
- Data processing is most important step for data preparation in a manner which is less time consuming activity. Hadoop is an ideal solution which can be used for converting the unstructured data to structure format, perform cleansing and store in unified Hive structures.
- The digital oil field provides PPDM compliance models for ease of integration and portability post standardization into a Digitized Platform.
- The quality of data on the Digitized Platform is verified by the stakeholders.
- Complex analytics and event processing is performed to find the drilling patterns, infer the lithology content based on various parameters of the oil well logs. In turn provide adaptors to 3rd party interfaces for data interpretation.
- Integration with the BI services and Enterprise Application Integration service to third party agents for advanced analysis and dashboard generation.



TECHNICAL PROCESS FLOW

There are five stages depicted in below diagram. Stating the lifecycle of the data process in big data platform.

a. Data Capture Stage

Fetch the customer's E&P data from various well in different phases. Apache Flume can be used for capturing Oil Well log data embedded in a standard formats such as Logical ASCII Standard files, Seismic Data files etc. Sqoop can be used for capturing data from RDBMS structured production data.

b. Data Storage And Preparation Stage

The massive volume of relevant data is then stores on Hadoop distributed file systems. Hadoop streams can be used for invoking the data preparation, massaging and cleansing scripts. The data preparation jobs can convert the unstructured data to structure format, platform cleansing and store in unified Hive structures. The data Governance can be carried out by tools such as Oozie and Zookeeper taking the security features of Hadoop.

c. Data Aggregation Stage

This is most important step which will aggregate the data from Hive/HBase or any other NoSQL database, so that analysis can be carried out on the aggregates.

d. Data Analytics Stage

In this step, further analytics to find the drilling patterns, infer the lithology content based on various parameters from the oil well logs. This step can be performed on another analytical database or In-Memory database residing outside the hadoop ecosystem.

e. Data Visualization Stage

The output from the analytics can be integrated to DW/BI systems for generating dashboards and scorecards so that decision makers can visualize and interpret the data.

MAPREDUCE

MapReduce is a programming model for handling complex combination of several tasks and it was published by Google. It is a batch query processor and can run an adhoc query for whole dataset and get the results in a sensible manner which has to be transformative. It has two steps: 1.Map: Queries are divided into sub queries and allocated to several nodes in the distributed system and processed in parallel. 2.Reduce: Results are assembled and delivered.

DATABASE

Oracle has introduced the total solution for scope of enterprise which requires Big Data. Oracle Big Data Appliance is a tool to integrate optimized hardware and extensive software into Oracle Database 11g to endure the Big Data challenges.

CONCULSION

A Hadoop based Big Data Framework with Hive as a central Data warehouse layer is widely used to create dynamic and unified structures. We can easily execute predefined or ad-hoc on the Hive. This acts as a unified integrated layer that can be easily augmented with current BI stack. The salient features of Hadoop/Hive based solution with respect to Oil and Gas E&P data management are:

- Scalable architecture to analyze terabytes to petabytes of multi structured well log data.
- Massive parallel processing providing unified view of the data from multiple wells during its lifecycle-be it at the planning, operations or post-completion stage.
- Integrated KPI framework-for commercials, operations, health and safety execution, productions, etc.
- An extendable PPDM complaint data-model and Energestics standards to manage the data with a partner ecosystem.
- Comparative analytics & correlations with wells in similar geologic conditions to help decision making for drilling oil wells.

Oil and Gas domain Ontology for easy interpretation of scientific terminology.

REFERENCES

Analytics Magazine (Nov-Dec 2012 issue)-How Big Data is changing the Oil and Gas industry by Adam Farris-Austin, Texas http://www.analytics-magazine.org.

Apache Hadoop wiki – http://wiki.apache.org/hadoop.

An Oracle whitepaper, Jan 2012 "Oracle: Big Data for the enterprise".

Grobelnik, Marko. Big Data Tutorial http://videolectures.net/eswc2012_grobelnik_big_data

Hilbert, M, Priscila L. 2011. "The World's Technological Capacity to Store, Communicate and ComputeInformation". *Science* 332 (6025): 6065. DOI:10.1126/science.1200970. PMID 21310967

MIKE 2.0, Big Data Definition http://mike2.openmethodology.org/wiki/Big_Data_Definition

White, Tom, Hadoop: The Definitive Guide.O'Reilly Media, ISBN 978-1-4493-3877-0.

http://www.forbs.com/sites/tomgroenfeldt/2012/01/06/big-data-big-money-says-it-is-a-paradigm-buster/

http://www.emc.com/about/news/press/2011/2010628-01.html.

White TH. 2009. The Definitive Guide. 2009. 1st Edition. O'Reilly Media.