PRAIRIE VIEW A&M UNIVERSITY
COLLEGE OF ENGINEERING
PRAIRIE VIEW, TEXAS

A SCALABLE AND DISTRIBUTED SEISMIC DATA TOOLKIT ON BIG DATA ANALYTICS PLATFORM

A Thesis
Submitted to the Graduate School
In Partial Fulfillment of the Requirements for
The Degree of

MASTER OF SCIENCE
IN
COMPUTER SCIENCE

Submitted By:

CHAO CHEN

Certificate of Approval:

(Advisor's name) Chairman Student Advisory Committee	(Department Head's name) Department Head Computer Science Department
(Dean's name) Dean College of Engineering	(Graduate School Dean's name) Acting Dean Graduate School

December 2016

ABSTRACT

A Scalable and Distributed Seismic Data Toolkit on Big Data Analytics Platform

December 2016

Chao Chen, B.S., southwest university of science and technology;

Chair of Advisory Committee: Dr. Lei Huang

Seismic volume is a basic data structure that has been widely used in geophysical computing, as well as in big data analytics applications of petroleum industry. It has been well studied in High Performance Computing (HPC) platforms. However, a little study has been done in big data analytics platforms. In this paper, we present an efficient Distributed Seismic Data Analytics Toolkit that works on the popular Apache Spark big data analytics platform. The toolkit supports different ways of seismic volume data distributions, repartition, transposing, access, and data parallelism with a variety of parallel execution templates. This work studies the performance characteristics of these seismic volume data operations with different configurations, current status of this big data analytics cloud at PVAMU as well as our research plan for future work. The scalability and parallelism are the main characteristics of this platform.

ACKNOWLEDGEMENTS

I would like to thank my advisor, professors, colleagues and friends who helped me a lot in the past two years. I cannot make it through and finish my thesis without your kind help.

First I want to thank my advisor Dr. Lei Huang, for giving me the chance to work in CloudComputing Lab, guiding my studies and researches in cloud computing. The facilities of CloudComputing Lab is a luxury for a student who want to do research on big data platforms. I would also like to thank Dr. Lin Li and Dr. Frizell, for spending lots of time to refine my thesis. Special thanks to Mr Ted Clee, for sharing his deep knowledge of seismic data analysis, which benefit me a lot in understanding so many geophysics terminologies in the use cases.

Also, I really appreciate all my colleagues and classmates worked in the lab in the past two years, for sharing so many interesting things from different cultures, and, of course, for helping and encouraging each other in work.

Finally I want to say thank you to my friend Yuzhong Yan, for always being the guy help me out at so many tough time in my life. I have learned so much from you not only as an engineer but also as a kind and wise man. Best wishes to you and your lovely family.

TABLE OF CONTENTS

Page

LIST OF FIGURES

FIGURE Page

LIST OF TABLES

TABLE Page

CHAPTER 1

INTRODUCTION

Petroleum is a traditional industry where massive seismic data sets are acquired for exploration using land-based or marine surveys. Huge amount of seismic data has already been generated and processed for several decades in the industry, although there was no the big data concept at that time. High Performance Computing (HPC) has been heavily used in the industry to process the pre-stack seismic data in order to create 3D seismic property volumes for interpretation.

1.1 Background

Currently, reflection seismology (or seismic reflection) is the most important method used in the petroleum industry to estimate the properties of the Earth's subsurface from reflected seismic waves and explore geophysics using the principles of seismology. The complete processing flow of this method involves data acquisition, data processing, data interpretation and attributes analysis [4].

As shown in Figure 1.1, data acquisition is performed by seismic sources such as dynamite or air gun generate and spread out seismic waves, which are reflected back from encountered materials underground and then recorded by the receiving sensors. To get data-scientists-usable 2D/3D seismic dataset, the collected seismic wavelet needs to be pre-processed by data processing methods including deconvolution, common-midpoint (CMP) stacking and migration. After data processing, the seismic events are geometrically re-located in either space or time to the location the event occurred in subsurface and create a complete image of subsurface [4].

The goal of seismic data interpretation and attributes analysis is to locate the

potential petroleum reservoirs from processed seismic reflections map. It involves deep knowledge of seismic attributes, geophysics, and intensive collaborations between data scientists and geophysicists. This thesis focus on develop a scalable and distributed toolkit to facilitates seismic data attributes analytics.

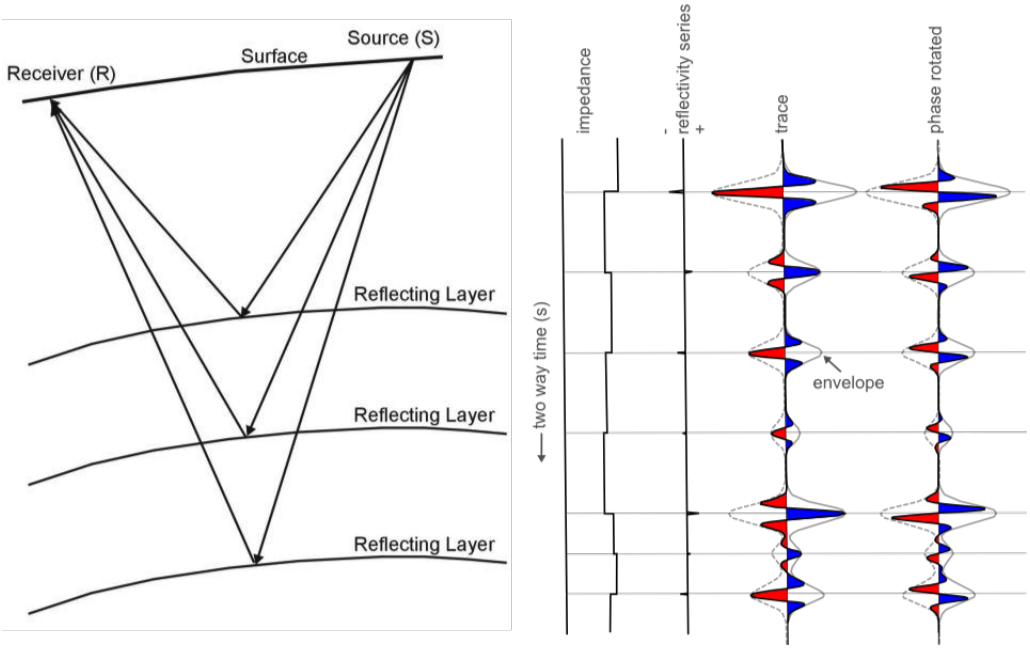

Figure 1.1: Reflection Seismology [5] [2]

1.2 Motivations

1.2.1 Big Data and Scalability

The emerging challenges in petroleum domain are the burst increase of the volume size of acquired data and high-speed streaming data from sensors in wells that need to be analyzed on time. Many types of captured data are used to create models and

images of the Earth's structure and layers 5,000-35,000 feet below the surface and to describe activities around the wells themselves, such as machinery performance, oil flow rates and pressures. With approximately one million wells currently producing oil and/or gas in the United States alone, and many more gauges monitoring performance, this dataset is growing daily [8]. Moreover, not only the real world datasets, but also the dimensions and complexity of datasets themselves increase dramatically, such as 4D even 5D and high density seismic data. The traditional HPC solutions are able to improve the performance of many computation-intensive models, however, most processing models are also data-intensive which are still the bottleneck for the whole workflow.

In many the data- and technology-driven industries, big data analytics platforms and cloud computing technologies have made great progress in recent years toward meeting the requirements of exploring the valuable information from fast-growing data volumes and varieties. Hadoop and Spark are currently the most popular open source big data platforms that provide scalable solutions to store and process big data, which deliver dynamic, elastic and scalable data storage and analytics solutions to tackle the challenges in the big data era. These platforms allow data scientists to explore massive datasets and extract valuable information with scalable performance. Many technologies advances in statistics, machine learning, NoSQL database, and in-memory computing from both industry and academia continue to stimulate new innovations in the data analytics field.

1.2.2 Complicated Workflow

Since the seismic data processing flow involves deep knowledges of geophysics, data science and computer science, it requests intensive collaborations among the scientists and developers from many different fields. This situation has long been

another bottleneck for the whole system. For an instance, most scientists are using MATLAB code to build their models, which is usually hard to translate to MPI codes. In most cases, the program needs to be reconstructed and parallelized by software engineers. For this situation, it is already a big challenge to both geoscientist and software engineers to understand each other's work, not to mention to maintenance or optimize the huge amount of legacy code in this industry.

1.3 Objectives

Geophysicists need an ease-to-use and scalable platform that allows them to advance the seismic data exploration process, design more intelligent algorithms to increase the drilling success rate. By Incorporating the latest big data analytics technology with the geoscience domain knowledge will speed up their innovations in the exploration/interpretation phase.

Although there are some big data analytics platforms available in the market, they are not widely deployed in the petroleum industry since there is a big gap between these platforms and the special needs of the industry. For example, the seismic data formats are not supported by any of these platforms, and the machine learning algorithms need to be integrated with geology and geophysics knowledge to make the findings meaningful.

The objectives of the work are to develop a seismic data analytics software development kits (SDK) to enable geophysicists to easily leverage the latest big data analytics technology to improve the seismic data exploration.

CHAPTER 2

RELATED WORK

The most famous definition of big data, comes from Gartner analyst Doug Laney, specifies the 3Vs characteristics: volume, velocity and variety [9]. By which volume means the amount of data, velocity stands for the real-time speed of data in and out, and variety is the range of data types and sources. As mentioned in previous chapter, the burst increase of volume size, high-speed real-time streaming data from sensors, and various types of structured, unstructured and semi-structured data coming from different stages of seismic data processing all together matches the 3Vs definition. It determines seismic data is costly to store, access and manage in traditional methodology. Therefore, new technology should be adopted to address these problems appropriately.

2.1 Researches in Petroleum Industry

Although lots of motivation exist in petroleum companies to adopt big data solutions to improve efficiency and reduce cost, only a few of them have deployed big data solutions. This situation may due to some technique barriers such as lack of technology knowledge, big data solution are not applicable in some steps of traditional workflow, and the cost and risk to convert legacy software to new platform etc. Moreover, there are lots of concerns of business-wise, such as the cost of infrastructure, data security issue (business or political restrictions on data accessing).

In [7], it mentioned working on the application of big data analytics in the Oil & Gas industry is in the experimental stage. Only a handful companies have adopted Big Data in the field:

1. Chevron proof-of-concept using Hadoop (IBM BigInsights) for seismic data processing;

2. Shell piloting Hadoop in Amazon Virtual Private Cloud (Amazon VPC) for seismic sensor data;

3. Cloudera Seismic Hadoop project combining Seismic Unix with Apache Hadoop;

4. PointCross Seismic Data Server and Drilling Data Server using Hadoop and NoSQL;

5. University of Stavanger data acquisition performance study using Hadoop.

Much of the software innovation that?s key to the digitization of big oil is happening at oil service contracting companies, such as Halliburton and Schlumberger, and big IT providers including Microsoft, IBM, Oracle and Open Source Projects [7].

2.2 Apache Hadoop and Spark

Since Google released its white paper series of big data processing technologies in 2004, the landscape of big data development has been changed profoundly. Many big data projects were inspired and developed based on MapReduce and Google File System framework. Hadoop and Spark is two most widely used open source big data solutions for many business and industry applications in recent years.

A year after the publication of MapReduce and Google File System framework, Doug Cutting and Mike Cafarella created an open source project Apache Hadoop, which has been used in many industries for a huge variety of tasks that all share the common theme of variety, volume and velocity of structured and unstructured

data [1]. The core of Apache Hadoop consists of a storage part, known as Hadoop Distributed File System (HDFS), and a processing part called MapReduce [3]. Distributed file system is fundamental to many main stream big data platforms as it is able to store data across number of storage devices of a cluster. Compare to traditional file system which holds sequential data on one device, HDFS provides far better scalability and support for parallel IO processing mode. Meanwhile, another open source big data project Apache Spark provides programmers with an application programming interface centered on a data structure called the resilient distributed dataset (RDD), a fault-tolerant collection of elements that can be operated on in parallel [6]. Since Spark itself does not provide distributed file system, it is usually installed on top of Hadoop, by which Spark could utilize HDFS interface to handle distributed data storage and access. The most different part between Hadoop and Spark is the parallel processing interface. MapReduce writes the result back to the storage after each reduction, while Spark utilizes RDD to handle most its operations and result in memory. This leads to up to100 times performance improvement compare to Hadoop in certain circumstances [6]. Moreover, Spark provides more advanced features such as real-time streaming processing interface and machine-learning library.

However, all of these big data platforms are designed for general purpose applications and focus on distributing the data, computation and IO overloads. Most MapReduce based framework do not have, or only have limited communication mechanisms between different maps, which is important to resolve the data and logical dependence problems in many complex applications. When it comes to the field of seismic data processing and analysis, the problem is more complicated. Since most scientists and researchers in petroleum industry do not have big data related knowledge or even computer science background, how to hide parallelism from them

and let them easily deploy their works on new platform is a big challenge for all the researchers.

CHAPTER 3

IMPLEMENTATION

The main goal of Seismic Data Analytics SDK is to develop a scalable and distributed software development tool to enable scalable computation and analytics of seismic volume datasets. This chapter will present the software architecture and the main functionalities of SDK, as well as some utilities we have built for user to deploy their applications on this big data platform.

3.1 Architecture

Seismic Data Analytics SDK is built upon Apache Hadoop and Spark. Figure 3.1 shows the software stack of a workable seismic data analytics platform. As shown in this diagram, the gray part is the OS layer, the elements with green color stands for the infrastructure layer of this big data platform, and on top of that, the components with blue color is the SDK. At the bottom of infrastructure layer, there is Hadoop Distributed File System (HDFS) that stores the big seismic data files by utilizing the large number of local disks. The Cassandra as a NoSQL database is also used to store seismic data, intermediate results and meta data. YARN and Mesos are used for resources management. Apache Spark is the data distribution and parallel execution engine based on the innovative idea of Resilient Distributed DataSets (RDD) concept. MLLib is included in the Spark as the machine learning package to enable machine learning based data analytics algorithms. OpenCV is the widely used image processing package that is used to provide image processing capability. Breeze is the numerical processing package including linear algebra, signal processing, statistics, and other numerical computation and optimizations written in Scala.

We have developed the seismic data RDD on top of Spark as the base distributed seismic datasets to enable parallel operations and machine learning algorithms. Geophysicists and data scientists can use the Seismic Data Analytics SDK to develop their own algorithms and leverage the capability of Apache Spark provides, as well as image processing, numerical computation, and deep learning packages.

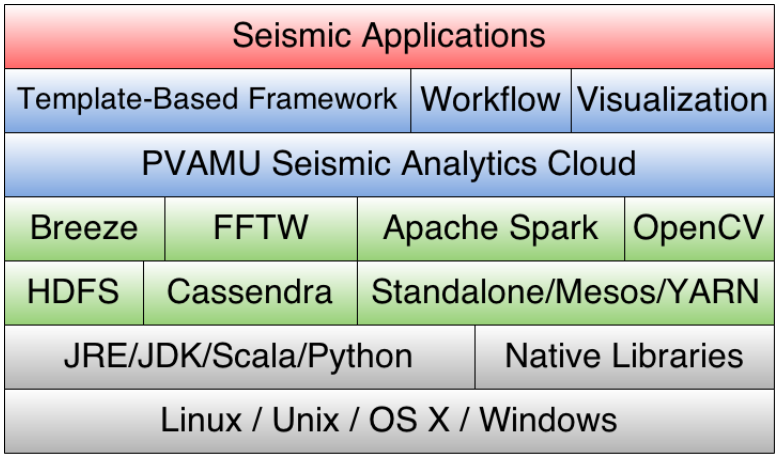

Figure 3.1: Software Stack of Seismic Data Analytics Platform

Figure 3.2 simplifies the development efforts for scalable and distributed computing and analytics of seismic datasets. It is built on top of the Apache Hadoop and Spark. The Hadoop provides a distributed file system(HDFS) and resource management system (YARN and Mesos), while Spark provides a high-level distributed data representation via Resilient Data Sets (RDD) and a data-parallelism execution engine. Seismic Data Analytics SDK provides configurable data distribution fashions for seismic volume data, as well as a configurable parallel execution interface to simplify the parallel programming efforts. Based on the functionality of SDK, we developed two useful utilities, parallel templates and data server, to facilitate SDK

for users to easily deploy their applications. Moreover, since Hadoop and Spark provide faults tolerance and task scheduling utilities, the toolkit inherits from them to provide fault tolerance and dynamic task scheduling for better reliability and task management.

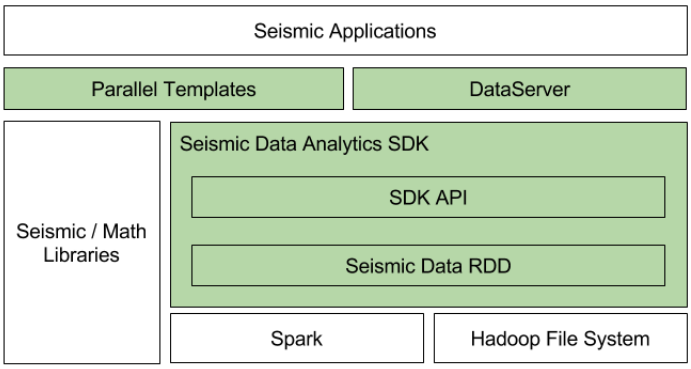

Figure 3.2: Framework of Seismic Data Analytics SDK

CHAPTER 4

EXPERIMENTS, RESULTS AND ANALYSIS

CHAPTER 5

CONCLUSIONS AND FUTURE WORK

5.1 Conclusions

In this paper, we focus on giving a friendly and high-efficiency solution to overcome the challenge of processing big seismic data. SAC was developed and verified with several typical applications, in which there is no prerequisite that users have parallel computing knowledge, and only the core algorithms need to be filled with the help of template provided by SAC. Such template provides friendly user interface for geophysicists without expense of performance. Some deep analysis about data partition, memory and network utilization are also given in this paper, which is a good experience for profiling any parallel program. From experiments and analysis results, SAC could handle big seismic data efficiently and easily.

5.2 Future Work

Although SAC has been proved to be a good candidate for processing seismic data, there are still some other work to improve. Current templates could hand the basic applications, but for some complicate cases, more templates need to be defined. If an application has many reduce actions such as Jacobi stencil codes that could not run in pipeline in whole scope, the performance could not boost too much in parallel. It is still a challenge for defining template if worker thread need to communicate with each other, and we will follow Spark community closely to find a good solution for sharing data between workers. In the future work, we will add more templates to make them fit more applications. More high level machine learning algorithms should be added to SAC and be applied to more advance seismic models. To make SAC

easy to be used by high level user and improve communication efficiency, workflow that could connect small piece of algorithms and Notebook for interactive seismic data processing are under development. Current visualization of seismic data in web interface is still in 2D mode, 3D view mode with remote rendering is already evaluated. For the performance optimization of parallel program in SAC, more deep research jobs are planned, such as adjusting GC parameters, GPU optimization and OpenSHMEM etc. In the view of applications developers, more data and computing models are need to investigate, such as streaming data, hybrid mode integrating with legacy codes etc. Fortunately, we had setup good relationship with many oil & gas companies and related service companies through this project, and their requirements in applications will provide some right directions for our research. In summary, there is still a long way to go for solving big seismic data processing problems, but we are already on the way.

REFERENCES

[1] 5-google-projects-changed-big-data-forever.

[2] E is for envelope. http://agilegeoscience.squarespace.com/journal/2011/3/23/e-is-for-envelope.html. [Retrieved: July, 2015].

[3] Hadoop Introduction. http://hadoop.apache.org/. [Retrieved: January, 2014].

[4] Reflection seismology. http://en.wikipedia.org/wiki/Reflection_seismology. [Retrieved: June, 2015].

[5] Seismic Reflection Methods. http://www.epa.gov/esd/cmb/GeophysicsWebsite/pages/reference/methods/Surface_Geophysical_Methods/Seismic_Methods/Seismic_Reflection_Methods.htm. [Retrieved: July, 2015].

[6] Spark Lightning-fast cluster computing. http://spark.incubator.apache.org/. [Retrieved: January, 2014].

[7] Baaziz A. How big data is changing the oil and gas industry, December 2013.

[8] Adam Farris. How big data is changing the oil and gas industry. http://analytics-magazine.org/how-big-data-is-changing-the-oil-a-gas-industry/, November 2012.

[9] Andrea De Mauro, Marco Greco, and Michele Grimaldi. A formal definition of big data based on its essential features. *Library Review*, 65(3):122–135, 04 2016.