Article can be accessed online at http://www.publishingindia.com

CDPSM: A New Optimized Progressive Big Data Analytics For Partial Cancer Data using Amazon EMR

Shyam Mohan J. S.*

Abstract

Identifying of symptoms and treating cancer requires a thorough investigation and research requiring analysis of multiple levels available (partial or full) cancer data. Cancer data is spread across multiple data sources and data warehouses which are decentralized and are in different locations. Therefore only half or partial data is available. Progressive analytics provide an efficient way for querying data from various data clusters where each cluster contains only a piece of the examined data. We propose an effective framework to perform analytics over the available cancer data say Cancer Data Progressive Sampling Model (CDPSM) built for partially available cancer data deployed on Amazon EMR. Through a large number of experiments, we reveal the advantages of the proposed model and give numerical results comparing them with a deterministic model. These results indicate that the proposed model can efficiently reduce the time for performing progressive data analytics over partial cancer data and maintaining the quality of the result at high levels.

Keyword: Big Data, Progressive Sampling

Introduction

The process of collecting, organizing and analyzing the data collected from various application domains like financial services, life sciences, mobile services, etc. is known as Big Data as most of the data is unstructured. The main aim of performing analytics is to discover patterns from hidden data sets and to provide meaningful information for effective decision making. Effective decision making will be successful by data driven by analytics-generated insights. Majority of the analytics are concerned with batch processing systems built on top of the Hadoop. Computing systems for big data generally fall into two major categories with regards to time constraint. They are:

1. Batch processing, in which large volumes of on-disk data with no time constraints (e.g., MapReduce and GraphLab) is analyzed.
2. In-memory streaming processing, where the data is analyzed in real-time or short period of time (e.g., Storm, SAMOA). Huang and Liu proposed that next-generation computing systems for big data analytics should be capable of providing good hardware and software to match between big data algorithms and the underlying computing and storage resources.

Challenges and Problem Statement

Currently there are many accessible data sources which provide information relating to any gene viz., mRNA or protein sequence. mRNA is estimated by the number of known sequences which is called Sequence Retrieval System (SRS). Majority of the medical data sources maintained by different organizations is updated frequently. For example, Nucleotide or protein sequences with the same emphasis are updated at different intervals with various benchmarks and standards and majority of the databases are outdated and contain irrelevant information. One of the challenges for cancer data is that the information stored is decentralized and is growing exponentially. For testing or for diagnosis of cancer data is a difficult task as the data is continuously updated viz., the cataloging and assessment behavior of dynamic biological regulation is incomplete. New categorical discoveries and their related information have to be constantly and progressively built onto any comprehensive content structures. In our work, we built

* Assistant Professor, Sri Chandrasekharendra Saraswathi Viswa Mahavidyalaya, Kanchipuram, Tamil Nadu, India.
Email: jsshyammohan@kanchiuniv.ac.in

a tensor based framework over cancer data and perform progressive analytics from partial information obtained which is used for treating and diagnosis of cancer.

Background and Literature Survey

According to Human Genome Project estimates, the human genome DNA contains around 3.2 billion base of pairs distributed among twenty-three chromosomes translated to about a gigabyte of information. By adding gene data, X-ray and NMR spectroscopy data, the volume increases dramatically in gigabytes or petabytes.

Some of the data collected from various repository are shown in table 1 and also can be found in references.

Contribution

CDPSM- a New Progressive Analytics Model for partially available Cancer Data

A new progressive model for partially available cancer data is proposed called CDPSM (Cancer Data Progressive sampling model).Without loss of generality; we assume that the users can encode their own sampling data i.e., by dividing them into various tuples or clusters at various intervals. For managing cluster data, schedulers are responsible for executing queries where the data is split into number of pieces or clusters for query execution and thereby achieving data parallelism and effective query composition. By performing successive progressive analytics on samples, they get incrementally processed providing a significant performance benefit. Schedulers work on the statistical assumptions and hence don't require any user involvement. Introducing CDPSM into an existing relational engine is easy because majority of the data appears in text. Implementing it on unstructured data is a challenging task. The table below shows the input data (taken in numeric) with progressive intervals and we rely on partial data.

Table 2: Input Data with Progressive Intervals

Interval	*User*	*Ad*
(0,∞]	user 0	a0
[1,∞)	user 1	a1
[2,∞)	user 2	a2

Amazon EMR

For effective and quick processing of vast amounts of data, we use Amazon Elastic MapReduce (Amazon EMR) web service. Amazon EMR provides effective Hadoop framework for processing huge and vast amount of data and hence providing an easy, fast, and cost-effective method for dynamically scalable Amazon EC2 instances Amazon EMR effectively handles big data use cases, log analysis, etc.

Taking Amazon EMR to the Cloud with Hadoop MapR Distribution

The MapR Distribution for Hadoop makes it easy for provision and managing Hadoop in the AWS Cloud in Amazon Elastic MapReduce (Amazon EMR). MapR is used for real time handling of cancer data and when used across various health care organizations for performing suitable data analytics will lead to effective treatment and diagnosis of cancer data and thereby providing a best proven platform for Big Data platform. Fortune 100 and Web 2.0 companies have already started using MapR for their organizations.

Running Parallel Hadoop Jobs in CDPSM Amazon EMR Cluster Using AWS Data Pipeline

For every available partial input of cancer data, we set start and limit for the cluster data without loss of generality. The input is a multi-stage job generated by partially available cancer data or simply Hadoop jobs. Each job consists of partial input data files, a partitioning key (or mapper), and a progressive reducer. For progressive analytics, we take Stream Insight to process cluster data. The Hadoop jobs can be run in parallel in clusters using Amazon Web Service (AWS) pipeline in CDPSM Amazon Elastic MapReduce (Amazon EMR).Cluster utilization can be increased using EMR. A scheduler is run on the top of Amazon EMR clusters for monitoring Hadoop activities and is responsible for running Hadoop activities in the cluster by assigning them to specific queues. New data arriving during real time processing can be specified to core Amazon EMR nodes and are automatically assigned to EMR clusters.

Table 1: Data Collected from Various Repository

S. No.	Repository	Sequence or category of Data	Data (Apprx.)	Data Growth (Apprx.)
1.	Prism (Progressive sampling Model)	Encode progressive sampling strategy into the data by augmenting tuples with explicit progress intervals.	Suitable for progressive analytics on big data in the Cloud.	Uses pipelining techniques.
2.	Now	Uses MapReduce computation paradigm	Binary Large Object Data (Blob)	Batch Processing
3.	GenBank (As on December 2014)	Nucleic acid sequences	178 million	Doubling in size for every 15 months
4.	SWISS-PROT database	Protein sequences	18 million	Doubling in size for every 15 months
5.	InSiteOne	Offers data archiving, storage, and disaster-recovery solutions to the health-care industry in U.S.	4 billion medical images and 60 million clinical studies from 800 clinical sites	Increasing at an approximate rate of about 12% per year.
6.	ESG (Enterprise Storage Group)	Forecasting Medical image data	Grows at a rate of 35 percent per year	2.6 million terabytes(2014)

Mechanism of Optimizing Time for CDPSM

For every available partial input of cancer data, we set start and limit for the cluster data. The major question is when to start and when to stop the model. The scheduler is responsible for retrieving partial results from clusters. Initially, we assume that the progressive interval starts from 0. Optimal Stopping Theory (OST) is used to stop CDPSM and to find the optimal or best time results based on sequentially observed random variables where stopping time is defined as a random variable $T \in 0, 1,...,\infty$.

The partial data is considered as random independent variables. The problem mentioned for

the partial data can be found in which are considered as a finite or an infinite horizon. For finite horizon scenario, the scheduler has to respond for a specific time interval and for an infinite horizon, the scheduler receives partial data and takes final decision in optimal time.

CDPSM For Multi Stage Processing

Schedulers

Effective resource allocation and job prioritization for a Hadoop cluster in CDPSM is provided by scheduler. Schedulers are chosen based on the type of application. For our stated problem, we choose Capacity scheduler and default scheduler both used interchangeably. Capacity scheduler uses queues for Hadoop clusters. The Capacity Scheduler is designed to run the jobs in Hadoop environment in a multi-tenant cluster which allows maximizing the throughput and allows sharing a large cluster. It sets limit for ensuring initialized or pending applications from the users and ensures stability of the cluster. Each job in CDPSM Hadoop is converted into a set of map and reduces tasks. CDPSM Hadoop resources in cluster enable sharing depending on the computing needs.

The configuration of capacity scheduler is done by the following commands:

HADOOP_CONF_DIR/capacity-scheduler.xml

HADOOP_YARN_HOME/bin/yarn rmadmin -refreshQueues

Launching an CDPSM on Amazon EMR cluster

CDPSM can be launched on the top of Amazon EMR cluster with MapR version 4.0.2 from AWS Management Console. It supports many editions of the MapR like Community Edition (M3), Enterprise Database Edition (M7), etc. The algorithm below shows launching an Amazon cluster.

Algorithm: Launching data instances in EMR cluster

Input:# of Mappers, X={a,b,c,….},Y={1,2,3,…}

Output: instances of cancer data

begin

Id =Partial Cancer Data Cluster;

Type =Data Cluster;

Hadoop Version =0.20;

Keypair =key value;

For each master Instance Type = k1.xlarge do

If (core Instance Type==k1.small) then continue

Core Instance Count= 30;

If (task Instance Type==k1.small) then continue

task Instance Count=30

Do boot strap Action set from D: //elasticmapreduce/bootstrap-actions/configure-hadoop,arg1,arg2,arg3, to D3: //elasticmapreduce/bootstrap-actions/configure-hadoop/configure-other-stuff,arg1,arg2;

End.

Evaluation

We have a multi-stage job generated by cancer data which is a partial data. Each job consists of partial input data files, a partitioning key (or mapper), and a progressive reducer as stated in section 6. For progressive analytics, each job consists of a special reducer which uses Stream Insight to process cluster data. Amazon EMR supports deployment on a cluster of machines. Due to this interesting feature, we have considered Amazon EMR for performing progressive analytics for partially available cancer data.

Experimental Setup

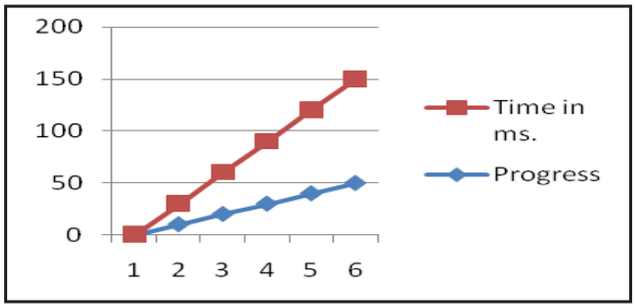
System Configuration

Machines in EMR are setup using Virtual Private Cloud (VPC) by changing the DNS Resolution and DNS hostname settings. Instances to communicate using EMR-managed security groups: The EC2 instance assigned is assumed to be default internal hostname. Input and output is stored in clusters. The cluster Id is known using VPC. Locally the system is of the configuration, 4GB RAM, and 1 TB of local storage, and 2Gbps allocated I/O bandwidth. We took 90 instances for our tests.

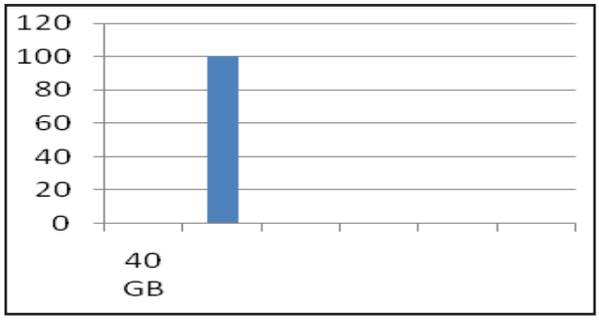
Datasets

We used the datasets available in for our evaluation based upon the aggregate amount of memory. We choose classification type as Brain cancer. Under this classification, we choose MicroRNA Data for Human Cancer data sets for performing analytics. Sample data sets are collected from over 299 patients that are available. We can even choose more samples based on the memory available. Factors in intergenic regions are also considered for data analysis. Input splits are created by shredding the data into various partitions with their corresponding Id. If the sample size is small we can rely on any algorithms like integer linear programming (ILP) algorithm. Stopping decisions are taken by Deterministic Stopping Model (DSM) which is considered for achieving the optimality. Missing or partial values of data can be found using parametric, non- parametric approach or Weibull distribution approach.

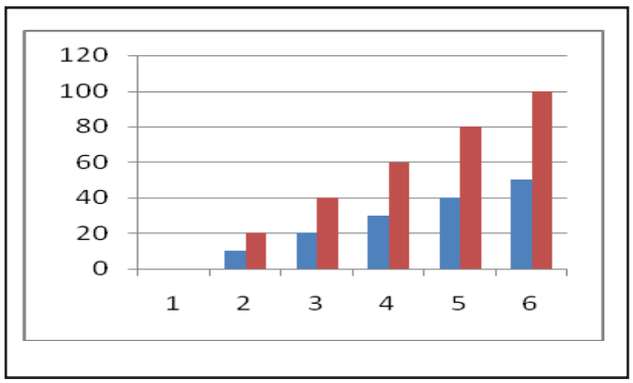
Figure a) shows the progressive computation for partially available cancer data (CDPSM) using Amazon EMR. Figure b) shows the performance of sample queries on Amazon EMR. Figure c) shows scalability of CDPSM for increasing data sets. Figure d) shows throughput of the machines.



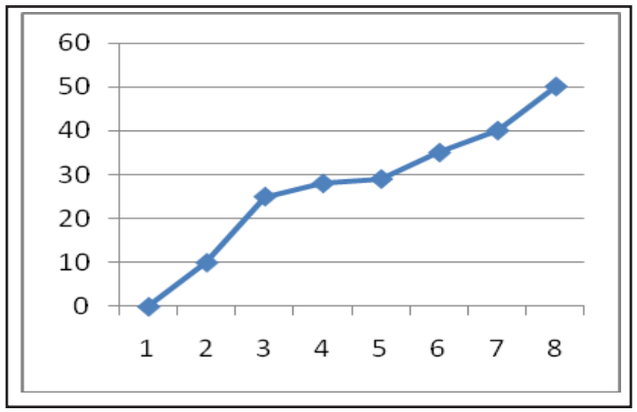
a) Progressive Computation-Time taken to process partial data in Amazon EMR.



b) Performance analysis of a sample query



c) Scalability with increase in data size



d) Throughput for number of machines

Fig. 1: Analysis of Various Data Sets on the Proposed Model

Conclusion

Progressive sampling on partial data is used to extract data for exploratory querying. Due to lack of proper tools for progressive analytics, obtaining results is a tedious task. We proposed a new progress model CDPSM deployed on Amazon EMR which allows efficient and deterministic query processing over partially collected sample data. For performing progressive sampling, we rely on Amazon EMR which provides a new framework for performing big data analytics over partially available cancer data where progress is achieved as a first-class citizen. Combination of all the above factors will lead to an effective and progressive big data analytics for partially available cancer data in optimized time.

References

Singh, S., & Singh, N. (2012). *Big data analytics,* In Proceedings of the International Conference on Communication, Information and Computing Technology.

Baldominos, A., Albacete, E., Saez, Y., & Isasi, P. (2014). *A scalable machine learning online service for big data real-time analysis.* In 2014 IEEE Symposium on Computational Intelligence in Big Data (CIBD), pp.1–8.

Huang, H., & Liu, H. (2014). *Big data machine learning and graph analytics: Current state and future challenges.* In 2014 IEEE International Conference on Big Data (Big Data), Oct. 2014, pp.16–17.

U.D. Energy, Insights learned from the human DNA sequence, what has been learned from analysis of the working draft sequence of the human genome? What is still unknown?, online, http://www.ornl.gov/hgmis, accessed on 2nd May 2011.

Hey, A. J., & Trefethen, A. E. (2003). The data deluge: An e-science perspective.

NCBI, Genebank statistics. Retrieved from http://www.ncbi.nlm.nih.gov/genbank/genbankstats.html, accessed on 2nd Aug. 2014.

Uniprotkb/swiss-prot protein knowledgebase release 2011_04 statistics. Retrieved from http://expasy.org/sprot/relnotes/relstat.html, accessed on 10th April 2011.

Uniprotkb/trembl protein knowledgebase release 2011_04 statistics, online, http://www.ebi.ac.uk/uniprot/TrEMBLstats, accessed on 10th April 2011.

Baluja, T. Electronic patient records will soon end doctor's scrawl on paper, the globe and mail. Retrieved from http://www.theglobeandmail.com/news/national/toronto/electronic-patient-records-will-soon-end-doctors-scrawl-on-paper/article1982647

Insiteone official website. Retrieved from http://www.insiteone.com/, accessed on 10th April 2011.

Nuclear Cardiology Markets, TriMark Publications, LLC, 2007.

EMR, E. (n.d.). News, Dell launches new cloud-based services for hospitals and physician practices, online. Retrieved from http://www.emrandhipaa.com/news/2011/02/21/dell-launches-new-cloud-based-services-for-hospitals-and-physician-practice, accessed on 3rd April 2011.

Efficient Machine Learning for Big Data: A Review, Omar Y.Al-Jarrah, Paul D.Yoob, Sami Muhaidat, George K. Karagiannidis,KamalTahaa, http://dx.doi.org/10.1016/j.bdr.2015.04.001 Elsevier-2214-5796/2015.

Chandramouli, B., Goldstein, J., & Quamar, A. (2013). *Scalable Progressive Analytics on Big Data in the Cloud.* In the Proceedings of the VLDB Endowment, 6(14). Riva del Garda, Trento, Italy.

Chaudhuri, S., Das, G., & Srivastava, U. (2004). Effective use of block-level sampling in statistics estimation, in: SIGMOD.

https://aws.amazon.com/elasticmapreduce/

Peskir, G., & Shiryaev, A. (2006). *Optimal Stopping and Free Boundary Problems,* ETH Zuerich, Birkhäuser.

Kolomvatsos, K., Anagnostopoulos, C., & Hadjiefthymiades, S. (2015). An Efficient Time Optimized Scheme for Progressive Analytics in Big Data. Retrieved fromhttp://dx.doi.org/10.1016/j.bdr.2015.02.0012214-5796, Big Data Research, February-2015, Elsevier.

https://hadoop.apache.org/docs/stable/hadoop-yarn/hadoop-yarn-site/CapacityScheduler.html.

http://www.broadinstitute.org/cgi-bin/cancer/datasets.cgi