Big Data Technologies: BigQuery vs MapReduce

Mohd Danish¹ Prof. Saoud Sarwar² Md Rashid Ashraf³

¹DuxConcept Technology, Jasola, New Delhi, India
²Department of Computer Science, Al-Falah School of Engg. & Tech. Faridabad, India
³BullsEyes Marketing Ltd, New Delhi, India

Abstract — The Big Data represents the incoming data that are continuously increasing in volume, variety and complexity. It is dynamic in nature and may or may not be stored in traditional DBMSs. In order to act effectively on big data, businesses must be able to understand data quickly, but also must be able to explore this data for value, allowing analysts to ask and iterate their business questions quickly. Querying massive datasets can be time consuming and expensive without the right hardware and infrastructure. Google handles the Big Data, coming from Gmail, Youtube and Google docs every second. Dremel is a query service that allows Google to run SQL-like queries against very, very large data sets and get accurate results in mere seconds. Dremel was launched to general availability, with its core feature, under the name of BigQuery. We will discuss how BigQuery compares to existing Big Data technologies like MapReduce and data warehouse solutions.

Keywords: Big data, BigQuery, MapReduce, Columnar Storage, data warehouse

I. INTRODUCTION

A. Big data

It is a collection of data sets so large and complex that it becomes difficult to process using on-hand database management tools or traditional data processing applications. The challenges include capture, curation, storage, search, sharing, transfer, analysis and visualization. [1,3] It can be defined in terms of 4 Vs :

1) Volume: The amount of data. Perhaps the characteristic most associated with big data, volume refers to the mass quantities of data that organizations are trying to harness to improve decision-making across the enterprise[2,3]. Data volumes continue to increase at an unprecedented rate[5].

2) Variety: Different types of data and data sources. Variety is about managing the complexity of multiple data types, including structured, semi-structured and unstructured data. Organizations need to integrate and analyze data from a complex array of both traditional and non-traditional information sources, from within and outside the enterprise. With the explosion of sensors, smart devices and social collaboration technologies, data is being generated in countless forms, including: text, web data, tweets, sensor data, audio, video, click streams, log files and more [1,2,5].

3) Velocity: Data in motion. The speed at which data is created, processed and analyzed continues to accelerate. Velocity impacts latency – the lag time between when data is created or captured, and when it is accessible. Today, data is continually being generated at a pace[2].

4) Veracity: Data uncertainty. Veracity refers to the level of reliability associated with certain types of data. Striving for high data quality is an important big data requirement and challenge, but even the best data cleansing methods cannot remove the inherent unpredictability of some data, like the weather, the economy, or a customer’s actual future buying decisions [7]. The need to acknowledge and plan for uncertainty is a dimension of big data [2].

B. Big data analytics

Big data analytics is the process of examining large amounts of data of a variety of types (big data) to uncover hidden patterns, unknown correlations and other useful information [3]. Big data analytics can be done with the software tools commonly used as part of advanced analytics disciplines such as predictive analytics and data mining. But the unstructured data sources used for big data analytics may not fit in traditional data warehouses.[3,5,6]

C. MapReduce

A variety of system architectures have been implemented for data-intensive and large-scale data analysis applications including parallel and distributed relational database management systems which have been available to run on shared nothing clusters of processing nodes including the MapReduce architecture [5,6]. The MapReduce programming model allows group aggregations in parallel over a cluster of machines. Programmers provide a Map function that processes input data and groups the data according to a key-value pair, and a Reduce function that performs aggregation by key-value on the output of the Map function.[4,7]

D. BigQuery

BigQuery is a cloud-based big data analytics web service for processing very large data sets. BigQuery was designed for analyzing data on the order of billions of rows, using a SQL-like syntax [11, 13, and 14]. There are 2 things that make BQ so fast: Columnar Storage and Tree Architecture.

1) Columnar Storage –

Instead of looking at the data is terms of rows, the data is stored as columns. The advantage of this kind of columnar storage is[10].-

i. Only the required values are scanned. This means that only 5-20 columns need to be accessed out of the thousands available [10,12]. This significantly reduces latency.

ii. Higher compression ratios are achieved when organizing data in a columnar format. Google reports that they can achieve columnar
compression ratios of 1:10 as opposed to 1:3 when storing data in a traditional row based format [16].

2) Tree Architecture –
This is used for processing queries and aggregating results across different nodes [15,17]. BQ is spread across thousands of servers. The data is shared across multiple machines. This helps retrieve data much faster.

E. Data Warehouse Solutions for OLAP/BI
Data warehouse solutions or appliances are being used for OLAP/BI use cases for many years. In OLAP/BI, we roughly have the following alternatives for increasing the performance of Big Data handling [22].

• Relational OLAP (ROLAP)
• Multidimensional OLAP (MOLAP)

1) Relational OLAP (ROLAP)
ROLAP is an OLAP solution based on relational databases (RDB). In order to make RDB faster, you always need to build indices before running OLAP queries. Without an index, the response will be very slow when running a query on Big Data.

2) Multidimensional OLAP (MOLAP)
MOLAP is an OLAP solution that is designed to build data cubes or data marts based on dimensions predefined during the design phase [22]. A weakness of MOLAP is that BI engineers must spend extensive time and money to design and build those data cubes or data marts before analysts can start using them [22,23].

II. COMPARISON: BIGQUERY, MAPREDUCE AND DATA WAREHOUSE SOLUTION
BigQuery, MapReduce and data warehouse are fundamentally different technologies and each has different use cases [19, 20, 21]. The following table compares these technologies and shows where they apply.

<table>
<thead>
<tr>
<th>#</th>
<th>Big Query</th>
<th>Map Reduce</th>
<th>Data warehouse solution</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>BigQuery is a Query service for the big data set</td>
<td>MapReduce provides programming model for processing large database.</td>
<td>Relational OLAP (ROLAP) - Multidimensional OLAP (MOLAP) are OLAP solutions based on relational databases (RDB) that always need to build indices before running OLAP queries.</td>
</tr>
<tr>
<td>2</td>
<td>It can process the unstructured data partially</td>
<td>Mapreduce can process unstructured data [18,20]</td>
<td>Multidimensional structured and semi-structured data</td>
</tr>
<tr>
<td>3</td>
<td>Easy to use for non-programmers (e.g. Analysts)</td>
<td>This requires the extra tools and efforts.</td>
<td>Knowledge worker (e.g., manager, executive, analyst)</td>
</tr>
<tr>
<td>4</td>
<td>Response time of BigQuery is very less</td>
<td>Time taken to respond, depends on the size of data (minute to days)</td>
<td>Extensive time and money to design and build those data cubes or data marts before analysts can start using them [20].</td>
</tr>
<tr>
<td>5</td>
<td>Ad hoc queries or trial-and-error data analysis</td>
<td>Batch processing of large dataset for time-consuming data conversion or aggregation.</td>
<td>ROLAP or MOLAP is not suitable for ad hoc queries or trial-and-error data analysis.</td>
</tr>
<tr>
<td>6</td>
<td>BigQuery is used for OLAP and Business Intelligence</td>
<td>MapReduce is the batch processing of large dataset for data conversion or aggregation.</td>
<td>ROLAP or MOLAP is used for OLAP and Business Intelligence</td>
</tr>
</tbody>
</table>

Table 1: MapReduce, BigQuery and Data warehouse appliances Comparison

The Authors [18, 19] have given an experimental result in figure2 which shows the result on 85 billion records and 3000 nodes. The result shows how fast the BigQuery (Dremel) is against the MapReduce model.

Fig. 1: MapReduce and Dremel (BigQuery Execution Time Comparison)

III. CONCLUSION:
BigQuery is a cloud based query service that processes huge datasets very quickly while MapReduce is suitable for long-running batch processes such as data mining. MapReduce is not suitable for ad hoc and trial-and-error data analysis. The turnaround time is too slow, and doesn’t allow programmers to perform iterative or one-shot analysis tasks on Big Data. When MR (Map-Reduce) jobs are used with column based storage it takes lesser time than the MR jobs with row based storage whereas BigQuery takes very few seconds for the same set of data. We can conclude that MapReduce is suitable for long-running batch processes such as data mining and BigQuery is the best choice for ad hoc OLAP/BI queries that require results as fast as possible. BigQuery and MapReduce compliments each other and BigQuery is the cost effective compared to traditional data warehouse solutions and appliances.

REFERENCES
Big Data Technologies: BigQuery vs MapReduce

IJSRD/Vol. 1/Issue 12/2014/002

All rights reserved by www.ijsrd.com


