What is BigQuery?

Sérgio Fernandes
Polytechnic Institute of Coimbra
ISEC – Coimbra Institute of Engineering
Rua Pedro Nunes, 3030-199 Coimbra, Portugal
Tel. ++351 239 790 200
sergio-d-fernandes@outlook.com

Jorge Bernardino
Polytechnic Institute of Coimbra - ISEC
Centre for Informatics and Systems of University Coimbra
Rua Pedro Nunes, 3030-199 Coimbra, Portugal
Tel. ++351 239 790 200
jorge@isec.pt

ABSTRACT
Big Data information is continuously increasing in volume and variety. This is the information that companies would like to quickly explore to identify strategic answers to the business. To overcome the problem of traditional database management systems to support large volumes of data arises Google BigQuery platform. This solution runs in the Cloud, SQL-like queries against massive quantities of data, providing real-time insights about the data. In this paper, we will analyze the main features of BigQuery that Google offers to manage large-scale data.

Categories and Subject Descriptors
H.2.4 [**Relational databases**]. H.2.6 [**Database machines**]. H.3.2 [**Information storage**].

General Terms
Design, Security, Standardization, Verification.

Keywords
BigQuery, DW, ETL, BigData, Google, Cloud, Database, SaaS.

1. INTRODUCTION
Nowadays, the amount of data being collected, stored and processed continues to grow rapidly. Therefore, high performance and scalability are two essentials requirements for data analytics systems. Querying massive datasets can be time consuming and expensive without the right hardware and infrastructure. Google BigQuery solves this problem by enabling super-fast, SQL-like queries against append-only tables, using the processing power of Google's infrastructure.

Google BigQuery is a cloud web service very attractive for its ease of use and functionality. It is ideal for businesses that cannot invest in infrastructure to process a huge amount of information. This platform allows store and retrieve large amounts of information in near real time with main focus on data analysis [1].

The remainder of this paper is organized as follows. Section 2 refers Cloud Computing and Section 3 overviews Big Data. Section 4 presents Google BigQuery platform and its main features. Section 5 presents the conclusions and suggests future work.

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the Owner/Author.
Copyright is held by the owner/author(s).
IDEAS'15, July 13–15, 2015, Yokohama, Japan.
ACM 978-1-4503-3414-3/15/07.
http://dx.doi.org/10.1145/2790755.2790797

2. CLOUD COMPUTING
Cloud computing allows access to computing resources easily scalable and virtualized via Internet. The use becomes simpler because users need not to have knowledge, experience or management of the infrastructure. There are usually three types of cloud services: Infrastructure as a service (IaaS) including servers, networking and storage; Platform as a service (PaaS) provides a higher level with custom applications and the third Software as a Service (SaaS) refers to one of the most known cloud services, which consists of applications running directly in the cloud provider [4].

3. BIG DATA
The Big Data is now an integral part of the business of a company helping in decision making. It offers several tools to process and analyze large amounts of information that are usually of Terabytes or Petabytes in size.

Hadoop is a landmark technology behind the Big Data allowing batch processing. The heart of Hadoop (2006) is the MapReduce (2004). However Google realized that MapReduce was not real-time solution to a large volume of data and to solve this problem developed Dremel application (2008). Dremel is designed to deliver exceptional performance on data distributed across multiple servers and SQL support.

In 2012 Google I/O event, it was announced the end of the Dremel and the beginning of BigQuery which became then the high-performance cloud service [1].

4. GOOGLE BIGQUERY
Google BigQuery platform is a SaaS as a model in the cloud. It is not a reporting system and does not have an interface that allows the operation of the information [1] but it is ideal to export results by Tableau, QlikView, Excel, among others including the tools of Business Intelligence (BI) as can be seen in Figure 1.

Projects are top-level containers that store information about billing and authorized users, and they contain BigQuery data. When we create a new project, it is identified by a name, authorized users and data [3].

4.1 Features
The solution presents some characteristics: Velocity, Scalability, Simplicity, Multiple permissions, Security, and Multiple access methods [1].

4.1.1 *Velocity*

BigQuery can process millions of information not indexed in seconds due to columnar storage, and tree architecture, as explained in the next sections.

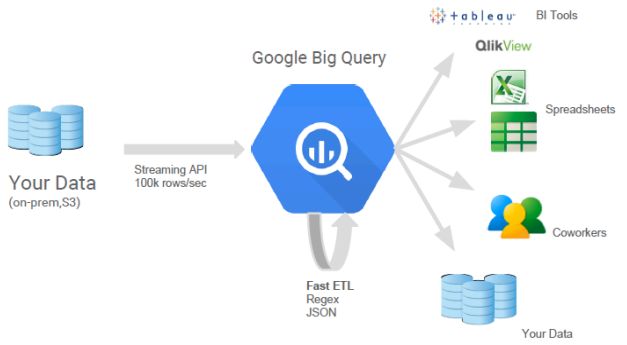

Figure 1 - Streaming Big Data [2]

4.1.1.1 *Columnar Storage*

The data instead of being stored in a line shape is stored as columns and thus storage will be oriented. In data analysis only the necessary columns are queried, largely reducing latency as is illustrated in Figure 2. This storage allows a higher compression ratio.

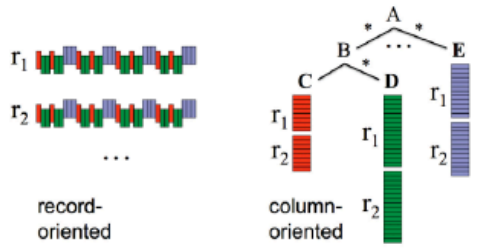

Figure 2 - Record vs Column Oriented Storage [1]

4.1.1.2 *Tree Architecture*

Used to query processing and aggregating the results between different nodes extended by thousands of servers using a binary tree. In this tree data is fragmented into multiple servers. This feature was born in Dremel.

4.1.2 *Scalablity*

It is the ability to manage huge data size with millions of records reaching terabytes of information, without space limits.

4.1.3 *Simplicity*

BigQuery provides a simple interface to upload and execute browse through a query language similar to SQL, as can be seen in Figure 3.

4.1.4 *Multiple permissions*

It is the capacity to manage different access permissions, read-only, editing or owner.

4.1.5 *Security*

To ensure security, the solution makes use of SSL (Secure Sockets Layer).

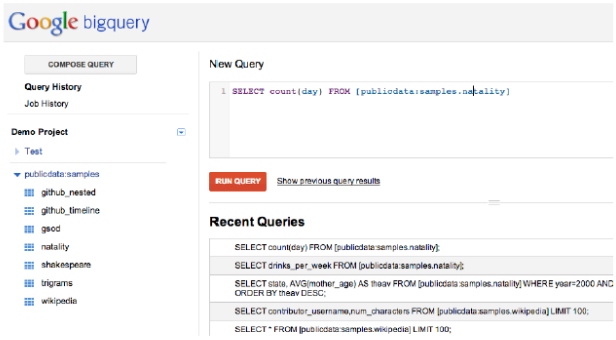

Figure 3 - Interface Simple [2]

4.1.6 *Multiple access methods*

We can access the service in different ways [3]. We can use a BigQuery Browser tool, a bq Command-line tool or a REST-based API.

4.1.6.1 *BigQuery Browser Tool*

With this tool it is possible to easily browse, create tables, run queries, and export data to Google Cloud Storage.

4.1.6.2 *bq Command-line Tool*

This Python command-line tool permits manage and query the data.

4.1.6.3 *REST API*

We can access BigQuery making calls to the REST API using a variety of client libraries such as Java, PHP or Python.

4.2 Cost

Many prices are practiced, existing free levels. It is charged to the customer by the total information processed but loading, reading or export information there is no associated cost. There is also budget option for a particular project [1].

5. CONCLUSIONS AND FUTURE WORK

In this paper we present Google BigQuery platform, which is a cloud-based database service that is able to process large data sets quickly. BigQuery allows to run SQL-like queries against multiple terabytes of data in a matter of seconds. It is the best choice for *ad hoc* OLAP/BI queries that require results as fast as possible. As a cloud-powered massively parallel query database it provides extremely high full-scan query performance and cost effectiveness compared to traditional data warehouse solutions and appliances. As future work we intend to test the solution with an assessment of their performance using real business data.

6. REFERENCES

[1] V. Villalba. Google BigQuery. Universidade Católica de Assuncíon, Paraguay.

[2] G. Mike. Hadoop, BigQuery, and Beyond. Enterprise Solutions Engineer – Google

[3] What is BigQuery? [online] https://cloud.google.com/bigquery/what-is-bigquery (Accessed 13 April of 2015)

[4] B. Carsten, K. Donald, K. Tim, L. Simon, "How is the Weather tomorrow? Towards a Benchmark for the Cloud", DBTest'09, June 29, 2009, Providence, Rhode Island, USA