Data Warehousing: A Preliminary Survey on Google

Ankita Singh¹ Prof. Sudhakar Jadhav²
¹MCA Student ²Professor
¹,²Department of MCA

1,2 Lokmanya Tilak College of Engineering, Navi Mumbai, Mumbai University

Abstract—Google, have one of the most powerful server networks in the known Universe. In terms of Google, space and computational capacity of databases.

Going with the Numbers
- 91 million searches per day
- Accounts for 50% of all internet searches
- Virtual profiles of increasing number of users

To access and store this huge data, Data warehousing architectures have been experiencing a rather dramatic evolution in recent years, and they will keep evolving into the foreseeable future. Scalable analysis on large data sets has been core to the functions of a number of teams at Google. Apart from ad hoc analysis of data and the creation of business intelligence dashboards by analysts across the company, a number of Google’s site features are also based on analyzing huge data sets, which range from simple reporting applications like Insights for the Google Advertisers, to more advanced kinds. So in this paper, we will study what data warehousing is and how it is shaping the Machine Learning, Business Intelligence, and Smart Advertising etc. Google is one of the most advanced company with advanced computational compatibility.

Key words: Data Warehouse, Business Intelligence, Mesa, Big Query

I. INTRODUCTION

A Data Warehouse (DW) is defined as, a subject-oriented, integrated, varies with time, nonvolatile collection of data in support of the management decision-making process. Data warehouses store stupendous amount of information from multi variant data sources which is used for query and analysis. That’s why, the data is stored in the multidimensional structure. A multidimensional model stores information into facts and dimensions. A fact contains the interesting concepts or measures (fact attributes) of a business process, whereas a dimension represents the perspective or view for analysing a fact (product, customer, time, etc.) using hierarchically organized dimension attributes.

A data warehouse is a database designed to enable business intelligence activities: it gives user’s platform to understand and enhance their organization's performance. And it has been designed for query analysis rather than for transaction processing, and usually contains data which has been derived from previous transaction data, and also include data from different sources. Data warehouses individually analysis workload from transaction workload and enable an organization to consolidate data from different sources. Which in turn helps in:
- Maintaining historical records
- Analysing the data to gain a better understanding of the business and to improve the business

Data warehouse process relational database, In addition, it can also include an extraction, transportation, transformation, and loading (ETL) solution, statistical analysis, reporting, data mining capabilities, client analysis tools, and other applications that manage the process of gathering data, transforming it into useful, actionable information, and delivering it to business users.

On an average, Google is subjected to 91 million searches per day, which accounts for close to 50% of all internet searching. Google stores every search a user and makes it into its databases. Collective search queries for over years’ worth of searches, total queries amount to more than 33 trillion database entries. Depending on the type of architecture of Google's databases, this figure could comprise ten to thirty petabytes of information.

Google is also in the business of collecting information on its users. Google combines the queries by individual user search with Google cookies, stored on users’ computers, which in turn can be used to create virtual profiles for each user (Which can be used to give user specific advertisements etc.).

<table>
<thead>
<tr>
<th>Activities</th>
<th>Duration</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of Emails Sent each Second</td>
<td>2.9 Million</td>
</tr>
<tr>
<td>Data Consumed by average households everyday</td>
<td>375 MB</td>
</tr>
<tr>
<td>Video uploads on YouTube per minute</td>
<td>20 Hours</td>
</tr>
<tr>
<td>Data processed per day in Google</td>
<td>24 PB</td>
</tr>
<tr>
<td>Data sent or received by mobile data usage</td>
<td>1.3 Exabyte’s</td>
</tr>
</tbody>
</table>

Table: Activities and their duration on Google

II. THE DATA WAREHOUSE ARCHITECTURE

Data warehouse architecture is being influenced by business practices and goals that continue to evolve. The reason: a well-aligned data warehouse reflects the business it serves.

Generally a data warehouse adopts a three-tier architecture. Following are the three tiers of the data warehouse architecture.

A. Bottom Tier

The bottom tier of the architecture is the data warehouse database server. Bottom tier is the relational database system. Which are used to pool, end tools and utilities to feed data into the bottom tier, which in turn simulate the Extract, Clean, Load, and refresh functions.

B. Middle Tier

Have the OLAP Server that can be implemented in either of the following ways.
- By Relational OLAP (ROLAP), which is an extended relational database management system. The Relational OLAP maps the operations on multidimensional data to standard relational operations.
By Multidimensional OLAP (MOLAP) model, which directly implements the multidimensional data and operations.

C. Top-Tier
This tier is the front-end client layer. The Top tier layer is in charge the query-reporting tools, analysis tools and data mining tools.

1) Data Cubes
This multi-dimensional data can be represented using a data cube as shown below. This figure shows a 3-Dimensional data Model.

Each cell represents the items sold of type ‘x’, in location ‘z’ during the quarter ‘y’. This is easily visualized as Dimensions are 3.

We can add more dimensions. This makes representing complex.

- Data cube is thus a n-dimensional data model

III. MESA: GOOGLE’S DATA WAREHOUSE
Google has a massive data warehousing system called Mesa. Mesa, a data warehousing platform designed to collect data from individuals to create virtual profiles, which in turn used to give internet advertising business. Advertising business on Google is stored in Mesa data warehouse, which is a distributed data warehouse that can manage petabytes of data while simultaneously delivering high scalability, redundancy, availability and fault tolerance. Mesa can update millions of rows per second, process billions of queries and retrieve trillions of rows per day. Mesa is a highly scalable analytic data warehousing system that stores critical measurement data related to Google’s Internet advertising business.

Google has found a way to stretch a data warehouse across multiple data centers, using Mesa architecture, its engineers developed that could pave the way for much more massive, extremely reliable and responsive cloud-based analysis systems.

Extending Mesa across multiple (geo_replicated) data centers allows the data warehouse to keep working even if one of the data centers fails.

Google runs an extensive advertising platform across multiple channels that serves billions of advertisements (or ads) every day to users all over the globe. Extended information attached with each served ad, such as the targeting criteria, number of clicks on ad number of impressions, etc., are recorded and processed in real time. This data is used extensively at Google for different use cases which including internal auditing, billing, reporting, analysis and forecasting.

The requirements for such a data store are:
- Atomic Updates.
- Consistency and Correctness.
- Availability.
- Near Real-Time Update Throughput.
- Query Performance.
- Scalability.
- Metadata and Online Data Transformation.

Mesa is Google’s solution to these technical and operational challenges. Even though subsets of these requirements are solved by existing data warehousing systems, Mesa is unique in solving all of these problems simultaneously for business critical data. Mesa is a distributed, replicated, and highly available data processing, storage, and query system for structured data.

A. Mesa Storage Subsystem
Each Mesa instance running in multiple Google data centers is completely independent and stores a full set of the advertising data. There are loads workers threads copying Mesa tables from one data center to another as each new facility is developed. When all tables are loaded onto the new Mesa cluster, update workers threads in the system apply changes to the data batches until it is in sync (or, more precisely roughly in sync because it is asynchronous) with the other Mesa instances. If one or more Mesa table(s) get corrupted in one facility (which is bound happen in some center) – the load workers threads are used to move a copy over from a known good version or previous good version in another datacenter.

Data in Mesa is continuously generated and is one of the largest and most valuable data sets at Google. Analysis queries on this data can range from simple queries such as, how many times a particular ad clicks were

---

All rights reserved by www.ijsrd.com 2011
performed for an advertiser on some particular day? to a more complex and elaborated query such as, How many ad clicks were performed for a particular advertiser, which matches the keyword ‘decaf’ during the last week of December between 7:00am and 12:00am that were though google.com for users in a specific geographic location using a particular device? Data from Mesa is inherently multidimensional, capturing all the minute facts and details about the overall performance of Google’s advertising platform in terms of various dimensions. These facts typically consist of two types of attributes: Dimensional attributes (which we call keys) and Measurable attributes (which we call values).

BigQuery is used for analyzing Big Data in the cloud. It makes SQL queries, run fast against petabytes of data in few seconds. Which is Scalable and requiring no setup or administration, BigQuery gives you real-time results based given data, which show some insight of data provided. BigQuery is Google’s fully managed, NoOps, data analytics service.

BigQuery is suitable for OLAP (On-line Analytical Processing) or BI (Business Intelligence) usage, where most of the queries are simple and done through a quick aggregation and filtering by a set of columns (dimensions).

D. Using BigQuery

- Finding particular records with some attached conditions. For example, you have to find requested logs with specified account ID.
- Quick aggregation of statistics with dynamically-changing conditions. For example, getting a summary of request traffic volume from the previous night for a web application and draw a graph from it.
- Trial-and-error data analysis. For example, detecting the cause of trouble and systematically collect values satisfying various conditions, by hour, day and etc…

V. CONCLUSION

The infrastructure of Google is scaling with the tremendous amount of data growth. By leveraging and developing a lot of open source technologies Google has been able to meet the demands placed on their infrastructure and they are working on many other enhancements to it in order to service those demands even more and in order to evolve this infrastructure to support new use cases and query patterns. The problem is that building software for a parallel system is more difficult than building it for a single all-powerful machine. The difficulty only increases as you break your application into tinier and tinier pieces and spread it across systems.

Real-time analysis over vast volumes of continuously generated data (informally, “Big Data”) has
emerged as an important challenge in the context of database and distributed systems research and practice.

In contrast, Mesa is a data warehouse that is truly cloud enabled (running on dynamically provisioned generic machines with no dependency on local disks), provides strong consistent and ordered version of the data. Mesa also supports petabyte-scale data sizes and large update and query workloads. BigQuery is the best choice for ad hoc OLAP/BI queries that require results as fast as possible. BigQuery is the cloud powered massively parallel query database that provides extremely high full-scan query performance and cost effectiveness compared to traditional data warehouse solutions and appliances.

REFERENCES
[10] https://cloud.google.com/bigquery/