### PROJECT REPORT



Topic: **Big Data Analysis** on Amazon Customer Review

By: GAURAV BHATT

College: Maharaja Surajmal Institute of Technology

## BIG DATA

Big data 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. Firms like Google, eBay, LinkedIn, and Facebook were built around big data from the beginning.

**4V’S OF BIGDATA:**

1.VOLUME: Data Quantity

2.VELOCITY: Data Speed

3.VARIETY: Data Type

4.VERACITY: Data Accuracy

**Why Big Data Analytics?**

**Cost reduction:** Big data technologies such as Hadoop and cloud-based analytics bring significant cost advantages when it comes to storing large amounts of data – plus they can identify more efficient ways of doing business.

**Faster, better decision making**: With the speed of Hadoop and in-memory analytics, combined with the ability to analyze new sources of data, businesses are able to analyze information immediately – and make decisions based on what they’ve learned.

**New products and services:** With the ability to gauge customer needs and satisfaction through analytics comes the power to give customers what they want. Davenport points out that with big data analytics, more companies are creating new products to meet customers’ needs.

**Hadoop**

Hadoop is an open-source framework to store and process Big Data in a distributed environment. Commercial distributions of Hadoop are currently offered by four primary vendors of big data platforms: Amazon Web Services (AWS), **Cloudera**, Hortonworks and MapR Technologies. Hadoop became a foundational data management platform for [big data analytics](https://searchbusinessanalytics.techtarget.com/definition/big-data-analytics) uses after it emerged in the mid-2000sIt contains two modules, one is MapReduce and another is Hadoop Distributed File System (HDFS).

* **MapReduce:** It is a parallel programming model for processing large amounts of structured, semi-structured, and unstructured data on large clusters of commodity hardware.
* **HDFS:** Hadoop Distributed File System is a part of Hadoop framework, used to store and process the datasets. It provides a fault-tolerant file system to run on commodity hardware.

**Big data analytics tools**

* HIVE
* PIG
* HBASE
* MAP REDUCE
* SCOOP

**HIVE**

Hive is a data warehouse infrastructure tool to process structured data in Hadoop. It resides on top of Hadoop to summarize Big Data and makes querying and analyzing easy. Provides SQL like query language named HIVE QL.

Initially Hive was developed by Facebook, later the Apache Software Foundation took it up and developed it further as an open source under the name Apache Hive. It is used by different companies. For example, Amazon uses it in Amazon Elastic MapReduce.

**Role of Hive**

* It provides ability to bring structure to various data formats.
* It provides simple interface for ad hoc querying, analyzing and summarizing large amounts of data.
* It allows access to files on various data stores such

as HDFS and HBase.

* Translates HiveQL statements into a set of MapReduce Jobs which are then executed on a Hadoop Cluster.

**PIG**

PIG is a platform for analyzing large data sets that consists of a high-level language for expressing data analysis programs, coupled with infrastructure for evaluating these programs.

* Pig is an abstraction on top of Hadoop
* Provides high level programming language designed for data processing
* Converted into MapReduce and executed on Hadoop Clusters
* Pig is widely accepted and used in Yahoo!, Twitter, Netflix, etc.

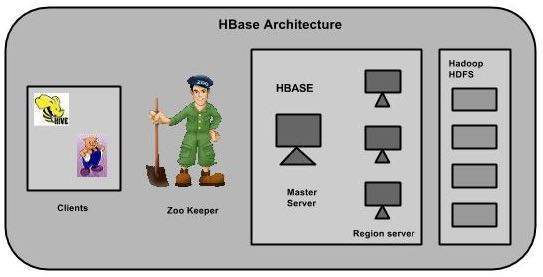
**PIG Features:**5

* **Ease of programming-** It is trivial to achieve parallel execution of simple, "embarrassingly parallel" data analysis tasks. Complex tasks comprised of multiple interrelated data transformations are explicitly encoded as data flow sequences, making them easy to write, understand, and maintain.

**Optimization opportunities-**The way in which tasks are encoded permits the system to optimize their execution automatically, allowing the user to focus on semantics rather than efficiency.

**HBASE:-**

* Hbase is a distributed column-oriented database built on top of the Hadoop file system. It is an open-source project and is horizontally scalable.
* One can store the data in HDFS either directly or through HBase.

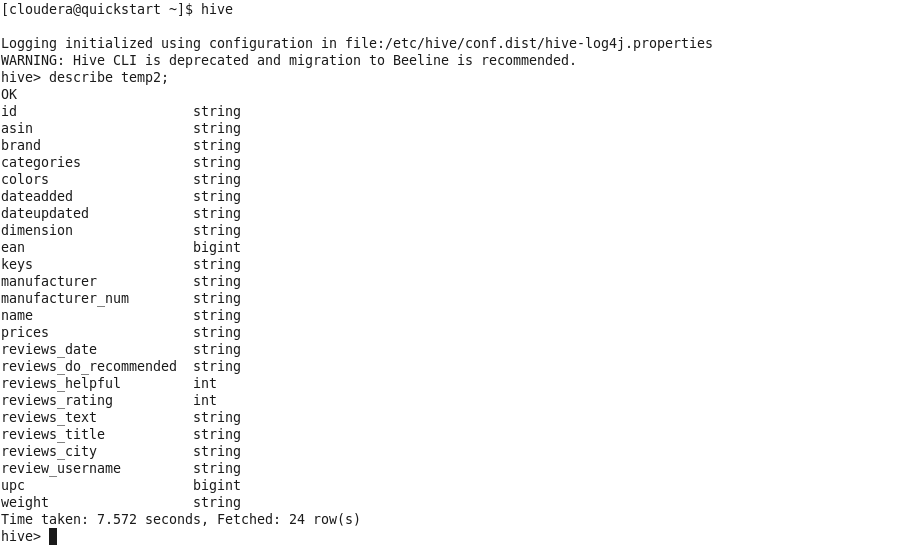


**Features of HBase-**

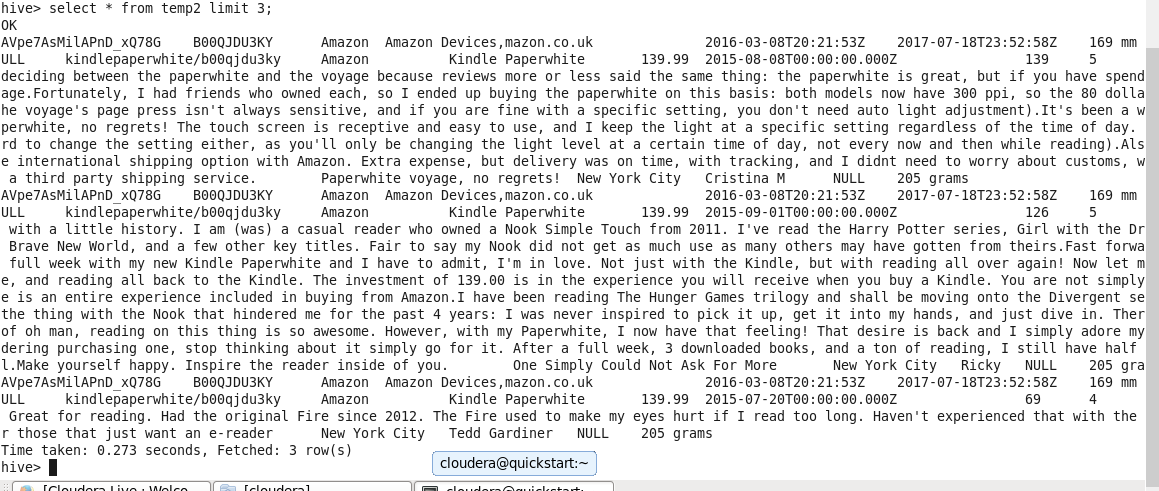
* HBase is linearly scalable.
* It has automatic failure support.
* It provides consistent read and writes.
* In HBase, tables are split into regions and are served by the region servers.

**Datasets: Amazon Customer Reviews**

**Data Description:**

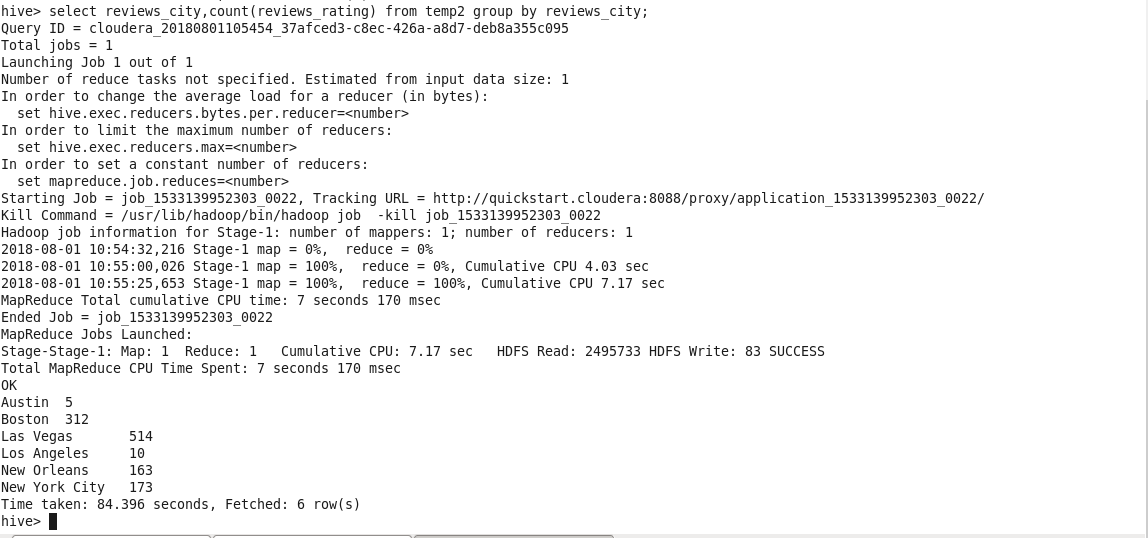
****

**Data Sample:**

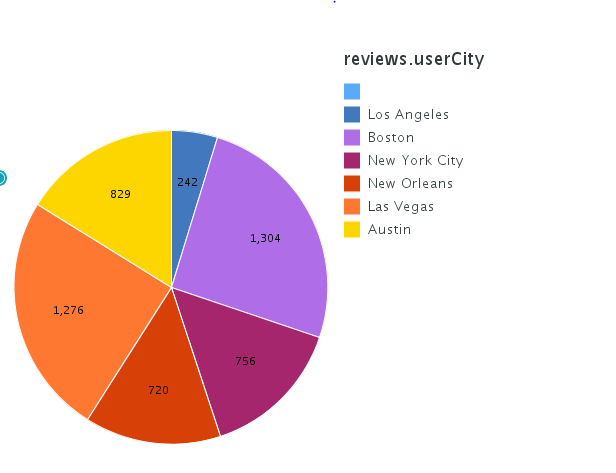
****

**Analysis on Amazon Customer Reviews**

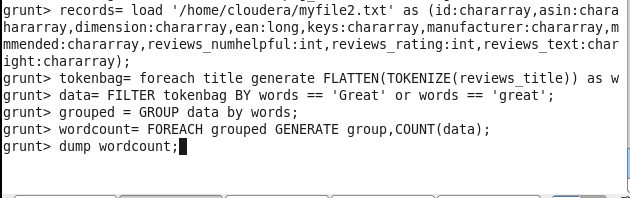
**Query 1:** Show the city from where maximum reviews are coming ?

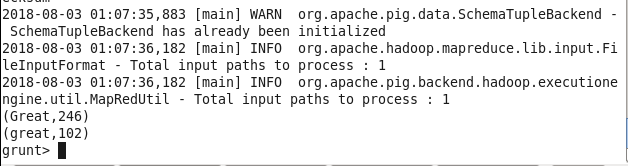


**VISUALIZATION by IBM COGNOS**

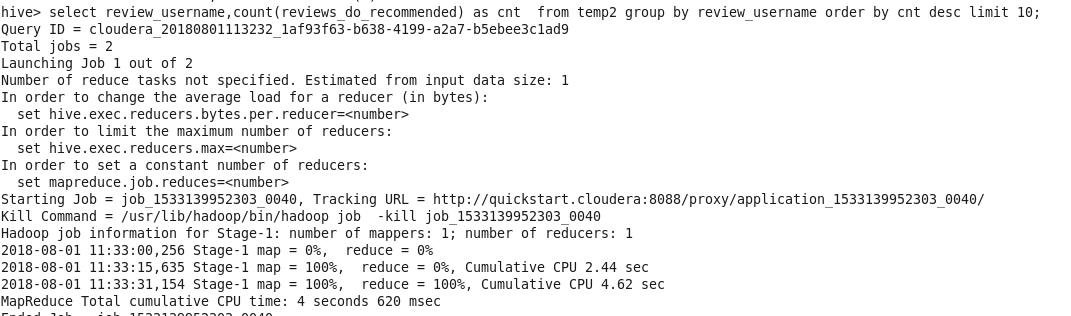
****

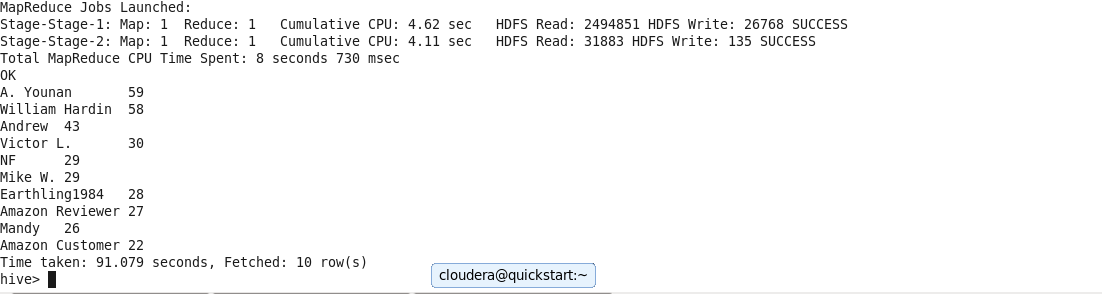
**Query-2:** Show the number of times the word ‘Great’ is used when customer is giving reviews.

****

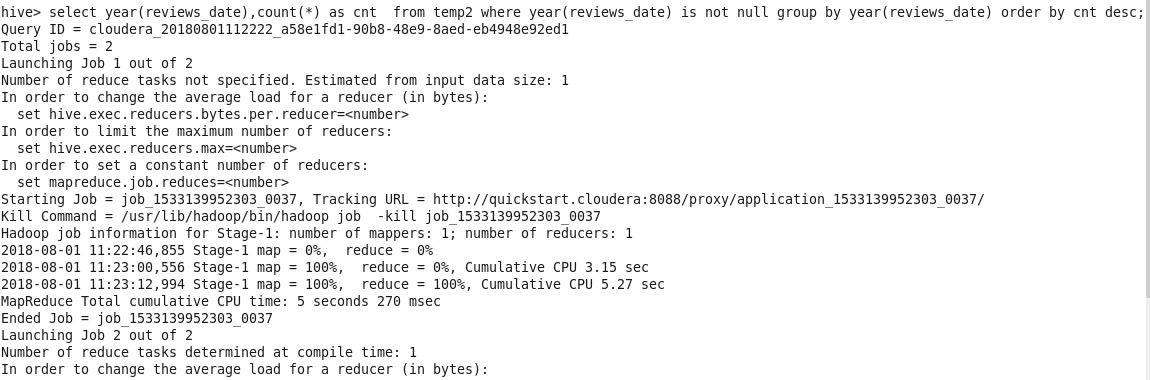
****

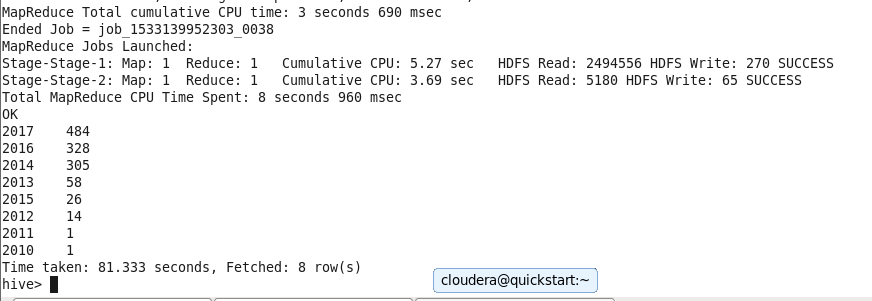
**Query-3:** Show the name of top 10 customers whose reviews are most recommended by others along with number of recommendation.



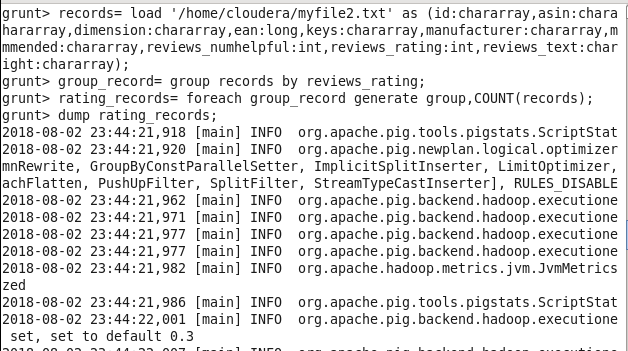


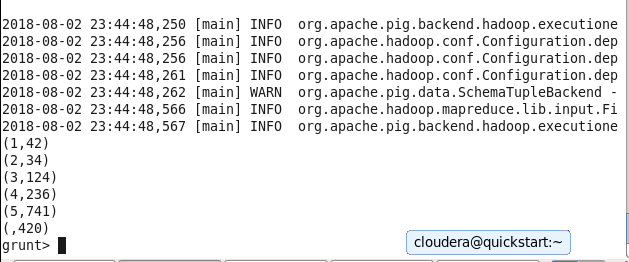
**Query-4:** Compare the number of product rated in various years.



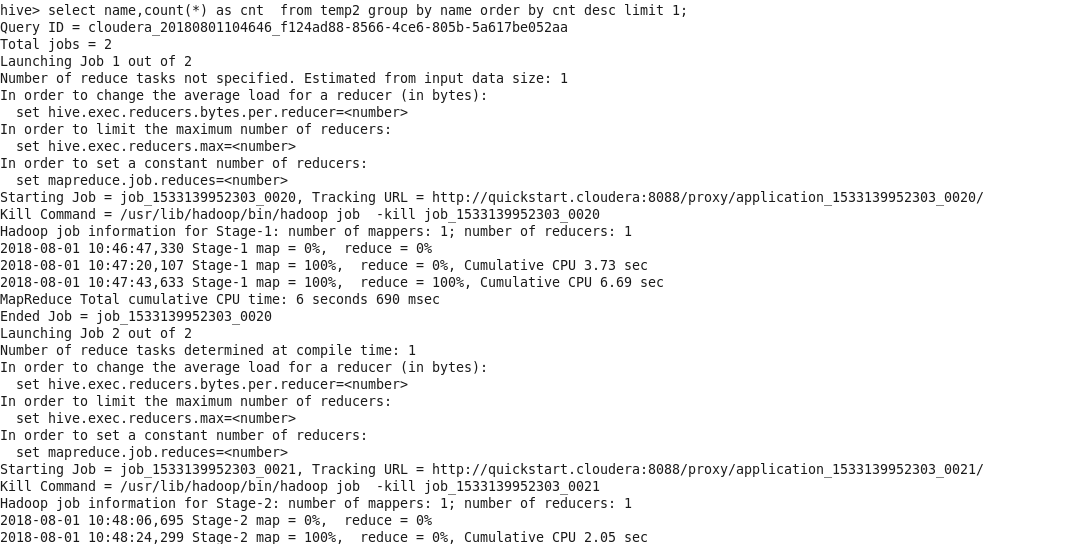


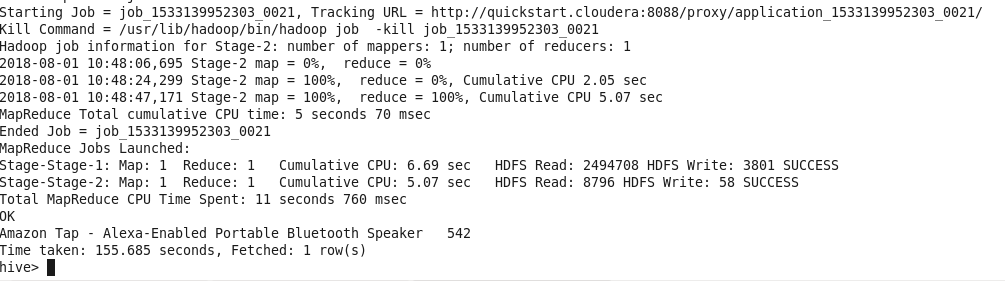
**Query-5:** Show the total number of rated products along with their rating.

****

****

**Query-6:** Which product is purchased by maximum number of customers ?

****

****

**THANK YOU**