**Reference Data Management**

**Confidential**

*Proprietary & Confidential*

**Copyright 2017 - 2018 INCEDO INC**. All Rights Reserved. No part of this document may be reproduced, stored in a retrieval system, distributed or transmitted, in any form, or by any means, electronic or otherwise, including photocopying, reprinting, or recording, for any purpose, without the express prior written permission of Incedo Inc.

**REQUESTS** For information requests or for obtaining permission for the use of this work, please submit a written request to: Marketing and Communications, Incedo Inc, Plot No. 248 Udyog Vihar, Phase IV, Gurgaon - 122015. Email: communications@incedoinc.com

**Document Revision History**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **S.No** | **VERSION NO** | **DATE** | **AUTHOR** | **APPROVAL BY** | **Remarks** |
| 1 | 0.1 | 31-01-2018 | Javeed Guruswamy Vahith Sasi Kumar | Incedo Hackathon Evaluation Team |  |
|  |  |  |  |  |  |

**Table of Contents**

[1. Abstract](#_Toc505269314)

[2. Solution Approach](#_Toc505269315)

[3. HIVE AND ORGANIZATION Data](#_Toc505269316)

[4. Design Architecture](#_Toc505269317)

[5. Business Impact](#_Toc505269318)

[6. Future Scope](#_Toc505269319)

# Abstract

The purpose of this solution is to apply Big Data Hadoop to financial analysis of **historical data of Financial Firms** in past years. Cloud computing, Hortonworks/Cloudera, is used for storing the historical data and for analyzing the data in Hive. The result contains its Growth by its highest volume for each industry (Basic, Consumer Durables, Technology, Energy, Transportation, Public Utilities, Consumer Non-Durables, Consumer Services, Capital Goods, Finance, Healthcare, and Miscellaneous). Additionally it has been found that the highest volume for the company Mega Capitalization. It is shown that financial data analysis can be done efficiently and easily using big data technologies like Hadoop and its ecosystem Hive by using cloud services like Hortonworks/Cloudera.

# Solution Approach

**Introduction:** Machine-generated data is growing exponentially from last several years. This data is generated in social networking sites via posts from many users, sensor data: to get climate information, purchase transaction records in large industry and many more. With the help of normal legacy systems, it becomes very difficult and expensive to store and analyze large scale data for data analyst. It is also time consuming process. This kind of large scale data with structured and unstructured format is called Big Data. However, Hadoop framework is growing now a days to store and analyze data and it is convenient for its functions. Hive is one of the ecosystems in Hadoop framework which is built by Facebook to analyze the data on Hadoop cluster. Hive syntax is based on SQL, so a person with the knowledge of SQL can easily work in Hive environment. The syntax used in Hive is called Hive QL (Hive Query Language).

Many companies have been using big data framework to analyze the data with relevant patterns among the data to target customer and market competition. We collect the data to keep in HDFS (Hadoop Distributed File Systems) and to analyze it using Hive on Hortonworks/Cloudera to find top companies by its highest revenue for each companies.

# HIVE AND ORGANIZATION Data

This section briefly describes Hive and Organization data set. Hive is a data warehouse infrastructure built on top of Hadoop for providing data summarization, query and analysis of organization data.

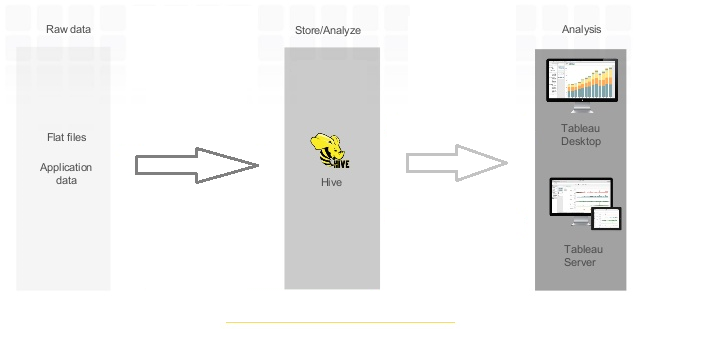
**3.1 Hive**

MapReduce codes are used to compute and analyze the Big data, But it requires high level expertise of Java, Ruby, Python and Perl. To ensure cost and time reduction Yahoo designed a new data flow language called Pig, which makes easy to analyze the data in 5-10 lines in Pig instead of 100 lines of Java code in MapReduce.

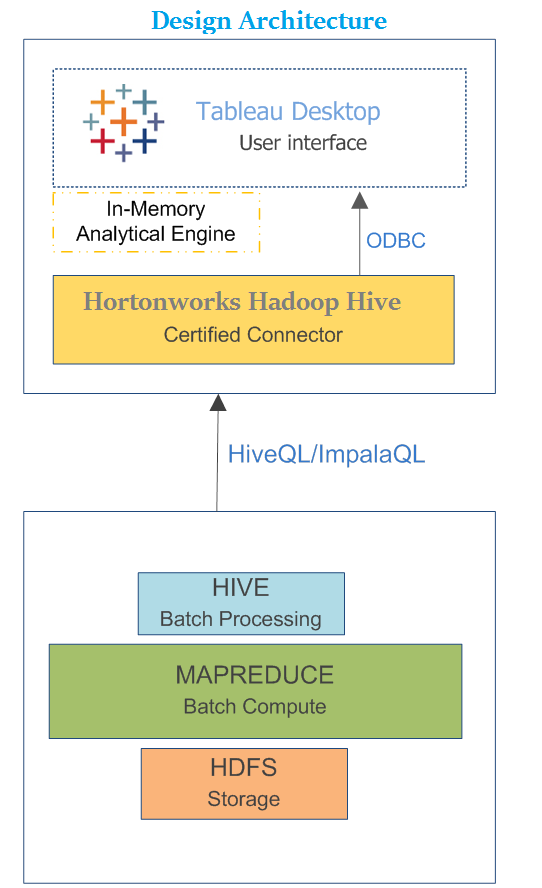
Hive and Hive QL platforms are designed by Facebook which has SQL like syntax. Hive runs on client machine and its queries are submitted to the Hadoop clusters on any local servers or cloud, which is transformed to MapReduce job. Based on our feasibility study, we have concluded to use Hive, Tableau and Hadoop clusters.

# Design Architecture

We come up with the below design architecture for this problem:



**Figure 1:** Data Analysis System



* The given dataset files loaded into temp path in HortonWorks then write the Hql Script with multiple constrains to execute the master table.
* We have multiple constrains for different data set, it containing data for all different companies we create a table using Hive QL on Hadoop system.
* Datasets having different structure, to make it align the structure used hive and after processing the data will be loaded into hive tables.
* Hortonworks Hadoop Hive connector used for fetch the data from hive tables and loaded into the Tableau server.
* Tableau used for the designing reports as we required.

# Business Impact

* This utility finds the companies whose financial metrics maximum from all other competitors.
* It has financial Reports of a single company with other competitors.
* They can forecast the growth of the company in future
* This makes analyzing and then visualizing your data a very arduous process.
* This can save huge amounts of time for repetitive analysis such as **Key Performance Indicator(KPI)** reports.
* Organization take right decision on right time using this Utility.

# Future Scope

* Client Attrition and Acquisition report Analyzing
* Dynamic parameters, where the options are updated from a data source.
* Better support for custom SQL