Rockwall Analytics
DATA & ANALYTICS ADVISORY

To: Sprocket Central Limited

From: Rockwall Analytics

Subject: Analytics, Information & Modelling - BIG DATA ANALYTICS

Date: December 18, 2020

THE OVERVIEW

Data Quality Assessment, Data Insights, Data Insights and Presentation

Data Quality
Assessment of data quality & completeness in preparation for analysis

INNOVATION & DIGITAL SOLUTIONS

Abstract

*Analytics, Information & Modelling helps organisations take the mystery out of **big data** and show them how to leverage their data resources to produce better business outcomes*

Rockwall Analytics's approach is based on the proposition that business success depends on what you actually do with your business information, not how much of it you control and collate

A 360 INFORMATION MINDSET

Abstract

in this information-driven age, leaders must take a 360 view of the extraordinary volume of data available – historic, current and predictive – so they can extract what they need and discover what they didn't know they need

with Rockwall Analytics's information-driven approach, we can give your organisation a holistic view of your data, enabling you to learn from and use it to make better business decisions, grow revenue, enhance operational capabilities, and manage enterprise risks and compliance mandates

STORY OF BIG DATA

In ancient times, people used to travel from one village to another village on a horse driven cart, but as the time passed, villages became towns and people spread out. The distance to travel from one town to the other town also increased. So, it became a problem to travel between towns, along with the luggage. Out of the blue, one smart fella suggested, we should groom and feed a horse more, to solve this challenge. When we look at this solution, it is not that bad, but do you think a horse can become an elephant? We don't think so. Another smart guy said, instead of 1 horse pulling the cart, let us have 4 horses to pull the same cart. What do you guys think of this solution? We think it is a fantastic solution. Now, people can travel longer distances in less time and even carry more luggage

The same concept applies on **big data**. Big Data says, we were okay with storing the data in our servers as the volume of the data is pretty limited, and the amount of time to process this data is okay. But, in this current technological world, the data is growing too fast and people are relying on the data a lot more. Also the speed at which the data is growing, it is becoming impossible to store the data into any server

BIG DATA DRIVING FACTORS

The quantity of data on planet earth is growing exponentially for many reasons. Various sources and our day to day activities generates huge volumns of data. With the invent of the web, the whole world has gone online, every single thing we do leaves a digital trace. With smart objects going online, the data growth rate has increased rapidly. The major sources of big data are:

- social media sites
- digital images/videos
- mobile phones
- purchase transaction records
- web logs
- medical records
- archives
- military surveillance
- eCommerce

and so on. All these sources of information amounts to around some Quintillion bytes of data. Currently, 2020, the data volume is around 40 Zettabytes which is equivalent to adding every single grain of sand on the planet multiplied by seventy-five

WHAT IS BIG DATA?

Big Data is a term used for a collection of datasets that are large and complex, which is difficult to store and process using available database management tools or traditional data processing applications. The challenge includes `capturing, curating, storing, searching, sharing, transferring, analyzing and visualization of this data`

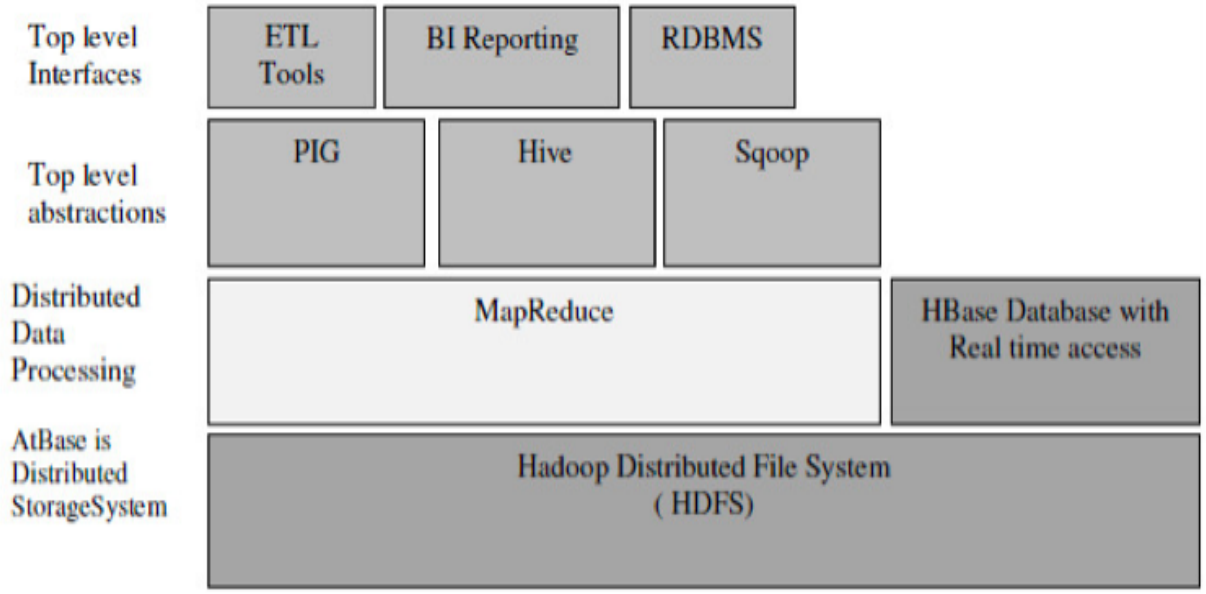
CHALLENGES PRESENT OPPORTUNITY

Big Data is emerging as an opportunity for organizations. Today, organizations have realized that they are getting lots of benefits from **Big Data Analytics**, as you can see in the image below. They are examining large datasets to uncover `hidden patterns, unknown correlations, market trends, customer preferences` and other useful business information

These analytical findings are helping firms in more effective marketing, new revenue opportunities, better customer service. They are improving operational efficiency, competitive advantages over rival organizations and other business benefits

HADOOP

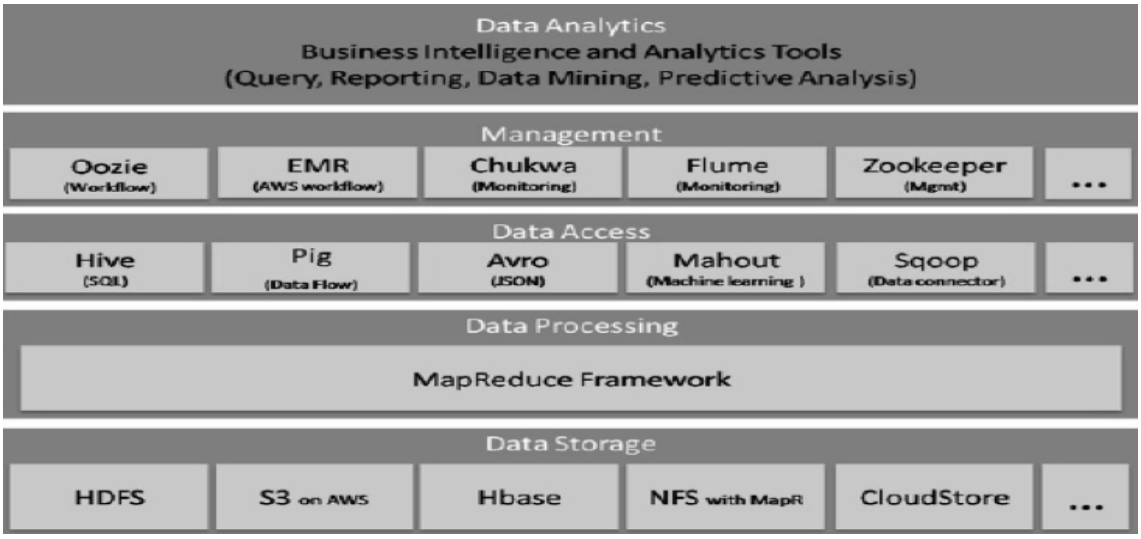
Hadoop is a collection of open-source software utilities that facilitate using a network of many computers to solve problems involving big data and computation. It provides a software framework for distributed storage and processing of big data using the MapReduce programming model



Big data analysis is the process of applying advanced analytics and visualizatio techniques to large datasets to uncover hidden patterns and unknown correlations for effective decision making. The analysis of Big Data involves multiple distinct phases which include data acquisition and recording, information extraction and cleaning, data integration, aggregation and representation, query processing, data modeling and analysis and Interpretation. This analysis covers:

- data quality assessment *information extraction and cleaning*
- data insights *query processing*
- data insights & presentation *analysis and Interpretation*

BIG DATA ANALYSIS TOOLS



THE BACKGROUND STORY

Sprocket Central Limited, (Sprocket) a medium size bikes & cycling accessories organisation, has approached Samira Variawa (Managing Partner) at Rockwall Analytics. Sprocket is keen to learn more about Rockwall Analytics's expertise in its Analytics, Information & Modelling

Samira discusses Rockwall Analytics's expertise in this space (you can read more here - README Link). In particular, she speaks about how the team can effectively analyse datasets to help Sprocket grow its business

Primarily, Sprocket needs help with its customer and transactions data. The organisation has a large dataset relating to its customers, but their team is unsure how to effectively analyse it to help optimise its marketing strategy

However, in order to support the analysis, Samira speak to Rihad the Lead Data Scientist - for some ideas and he advised that:

Abstract

... the importance of optimising the quality of customer datasets cannot be underestimated. The better the quality of the dataset, the better chance you will be able to use it to drive company growth

The client provided Samira with 3 datasets:

- Customer Demographic
- Customer Addresses
- Transactions data in the past 3 months

Samira decide to start the preliminary data exploration and identify ways to improve the quality of Sprocket's data

ADDITIONAL INFORMATION

Samira arrive at my desk after the initial client meeting. Samira has a voicemail on her phone which contains the following instructions

Abstract

[Voicemail transcript below]

Hi Samira,

Welcome again! Sprocket has asked you to assess the quality of their data; as well as make recommendations on ways to clean the underlying data and mitigate these issues. Can you please take a look at the datasets you've received and draft a memo to them identifying the data quality issues and how this may impact our analysis going forward?

I will send through an example of a typical data quality framework that can be used as a guide. Remember to consider the join keys between the tables too. Thanks again for your help

[Read email below]

Abstract

Hi Samira,

As per voicemail, please find the 3 datasets attached from Sprocket:

1. Customer Demographic
2. Customer Addresses
3. Transaction data in the past three months

Can you please review the data quality to ensure that it is ready for our analysis in phase two. Remember to take note of any assumptions or issues needed, as well as recommendations going forward to mitigate current data quality concerns

I've also attached a data quality framework as a guideline. Let me know if you have any questions

Thanks for your help

Kind Regards Sprocket point-of-contact

Our Task

Draft a memo to the Sprocket identifying the data quality issues and strategies to mitigate these issues. Refer to 'Data Quality Framework Table' and resources for criteria and dimensions which we should consider

Resources to help with the task

Data Quality Framework Table

Below is a list of the Data Quality dimensions you may use to evaluate the dataset. Some of these terms are common to the whole industry, so you may find more information and clarity on these terms by searching online