Big Data for a Big Decision

By

Elier Lara, Josh Moellman, Angelika Modawal

Submitted to
the Faculty of the School of Information Technology
in Partial Fulfillment of the Requirements for
the Degree of Bachelor of Science
in Information Technology

Elier Lara	Date 4/18/16
Josh Moellman	Date 4/18/16
Angelika Modawal	Date 4/18/16
Professor Jim Scott, Faculty Advisor	Date 4-18-16

University of Cincinnati
College of
Education, Criminal Justice, and Human Services

April 2016

Table of Contents

1. Abstract……1
2. Introduction……....2
3. Description…….....2
4. Problems Experienced……....3
5. User Profile…….…….4
6. Use Case Diagram…….....5
7. Proposed Budget……..…7
8. Project Timeline……8
9. Testing……...8
10. Conclusion……....9
10. Bibliography……....10

List of Illustrations

1. Figure 1…….…4
2. Figure 2…….…4
3. Figure 3…….…5
4. Figure 4…….…5
5. Figure 5…….…6
6. Figure 6…….…9
7. Figure 7……..9
8. Figure 8…….10

Abstract

College seniors often ask themselves where they should live after they graduate. They are often given anecdotal advice from friends and family or can consult the American Community Survey to see where recent graduates move. However, most college students do not use Big Data, based on personal criteria, to make the decision. The objective of this project was to provide a way to illustrate the power of the Hadoop architecture, defining the process of data ingestion and computation using analysis of US cities as a use case. The analysis covers information like weather, cost of living and job availability. The results will suggest the best possible location for recent college graduates to move to.

Introduction

Among the many decisions a recent college graduate may face, picking a career and place to live is among the biggest. We live in the age of big data – that is making decisions based on data, not emotion or feedback from others. As businesses are beginning to make the shift to making decisions based on data, so are tech-savvy college students. It makes sense for college students to want to understand how to leverage new technologies for their advantage. Our goal was to create a way to aid college graduates to make a more educated and well-rounded decision by using the newest upcoming technologies like Hadoop, Cloudera, and Tableau.

Description

Our project was sparked by the interesting concept of Big Data, which means taking vast amounts of data gathered anywhere and everywhere and analyzing them for patterns and trends. A common technology for storing big data is called Hadoop. Hadoop is an infrastructure designed to store and process large amounts of data. It specializes in unstructured data, which has no real structure or organization and therefore does not exist in a normal relational database. The infrastructure itself generally exists as one or multiple Hadoop "master" nodes that are connected to many "data" nodes. The data is then spread across all the nodes for storage and is then processed on the nodes themselves transmitting only the results back.

After looking at options for utilizing Hadoop to the fullest, and given the time constraint for the project, we felt it was best to use Cloudera. And, this product is a more automated way to process Big Data. Instead of manually installing all the pieces required, like Hadoop, Hue, Hive, and MapReduce, Cloudera installs all of those packages together in one instance.

So to carry out our vision for this project, we built a 9 node cluster consisting of one master node and 8 secondary nodes, otherwise known as “slave” nodes. This Cloudera-based infrastructure environment supported the data we formatted, imported, and queried to narrow down the optimal city choices for those who have just graduated.

Finally, to visually show the end result after all the data inputting and querying within Cloudera, we chose Tableau as our data visualization software. We feel adding this segment completes the project because it illustrates the data we are processing. In a sense, Tableau ties the analytics and infrastructure parts together into an end product that others can visually see through the graphs and charts produced.

User Profile

Application: Big Data for Big Decisions: Analyzing Big Data supported by a Hadoop Infrastructure
Potential Users: 1. Hadoop Engineers 2. Data Analysts 3. System Administrators 4. IT students
Software and Interface Experience: The end user for this should have experience in Linux command line, SQL querying, Business Intelligence tools, and databases. Also, knowledge in the languages Python or R would be helpful.
Experience with Similar Applications: 1. MapReduce 2. Yarn 3. SQL 4. Cloudera Manager
Task Experience: 1. SQL Querying 2. Linux Operating System and Command Line 3. Scripting languages – Python or R

4. Basic computer networking 5. VMware or VirtualBox if the infrastructure is virtual (ours is)
Frequency of Use: Anytime a graduating student wants to know where the optimal location for them to live after their college career, based on specific criteria.

Figure 1, User Profile

Use Case Diagrams

Hadoop Infrastructure

- The diagram below is illustrating how data will be spread across multiple nodes for faster processing. Now, we have more than 3 Slave nodes, but this diagram is just showing how data is processed

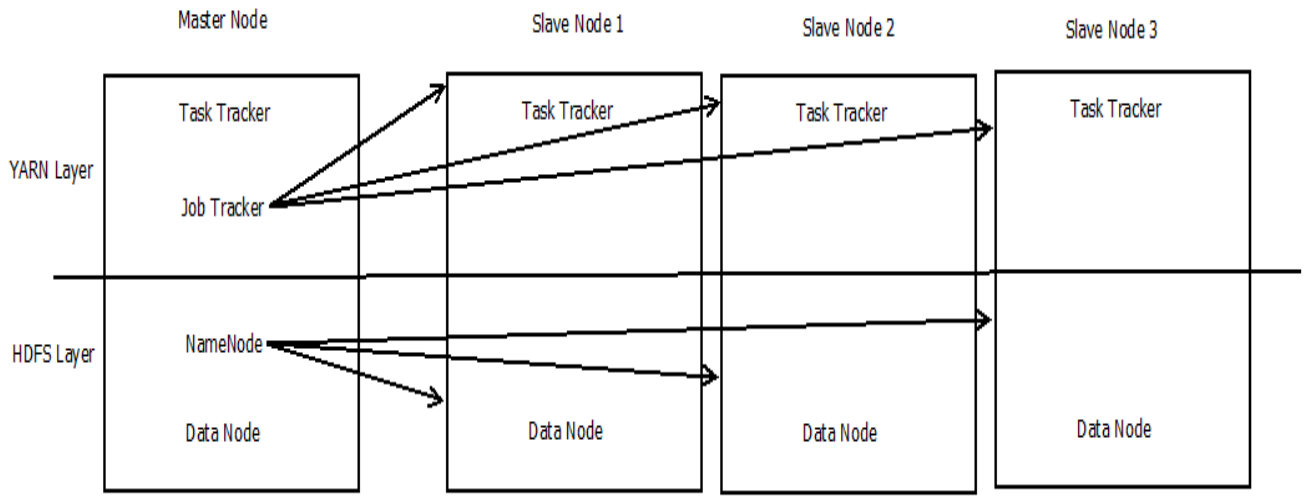

Figure 2, Use Case Diagram

Scenario Flowcharts

Scenario 1: User

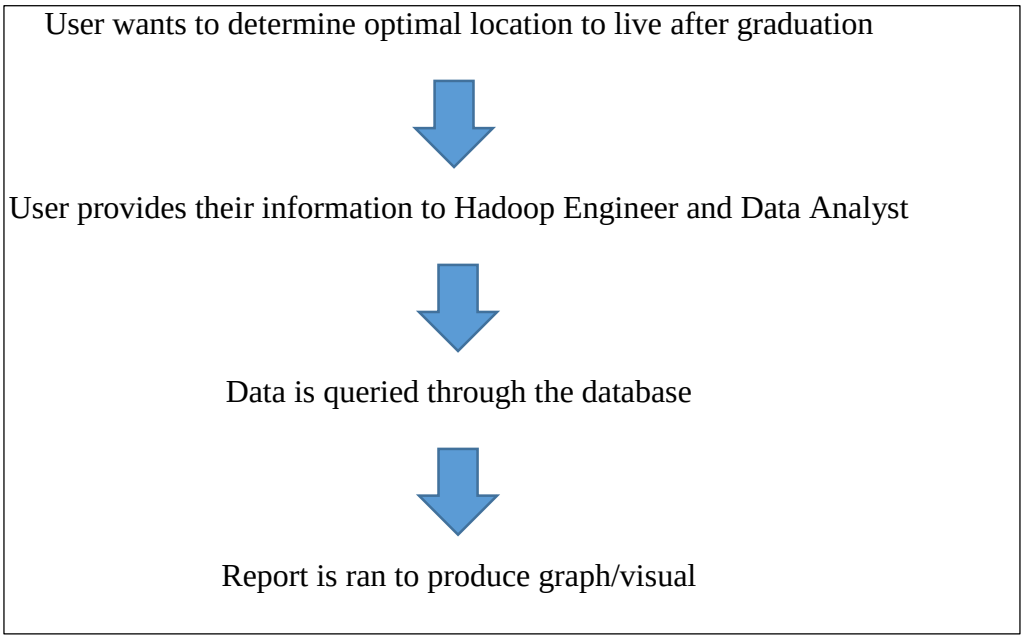

Figure 3, Scenario 1

Scenario 2: Engineer/ Analyst

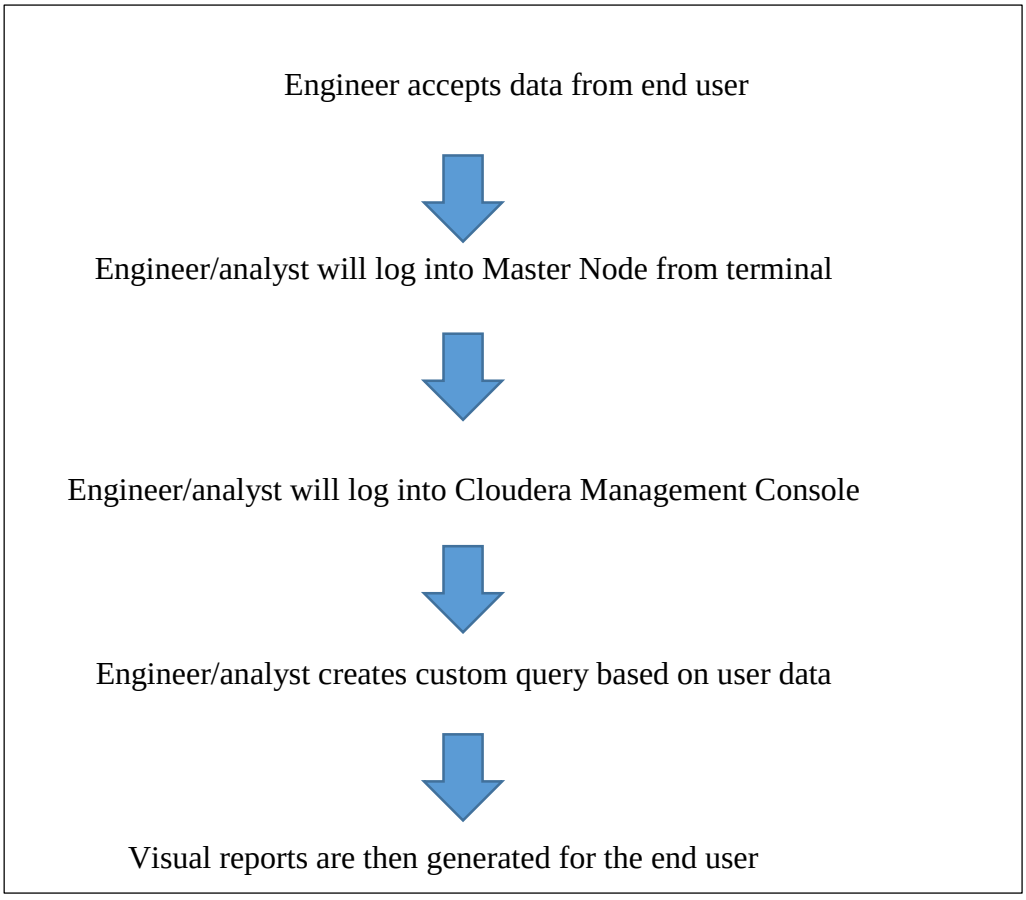

Figure 4, Scenario 2

Proposed Budget

The exact budget is an estimate based on current known costs. We did get a $35.00 grant from AWS Educate for cloud storage for our nodes. However, we may have to pay for additional resources as the project develops. Also, an Udemy course for learning Hadoop Analytics was purchased for $25.

Cost	Actual Cost	Business Cost
Udemy Course for learning Hadoop Analytics	$25	$25
Cloud Storage	$35.00 Amazon Grant	$100/monthly
Engineer Time	$0	$2000
VMware Workstation	$0	$250
CentOS	$0	$0

Figure 5, Budget

Project Timeline

Task Name	Duration	Start	Finish
Initial Research	**15 days**	**9/14/15**	**10/2/15**
Determine VM requirements	1 day	9/14/15	9/14/15
Initial Research of different Hadoop plugins	11 days	9/14/15	9/28/15
Solidify Criteria/Questions	2 days	9/28/15	9/29/15
Understand the data sources	2 days	9/28/15	9/29/15
Request space on the Sandbox	3 days	9/28/15	9/30/15
Understand how to import data into Hadoop	5 days	9/28/15	10/2/15
Extended Research	**20 days**	**10/5/15**	**10/30/15**
Research Cloudera Manager option	5 days	10/5/15	10/9/15
In depth research of Hue, Hive, and R project	4 days	10/12/15	10/15/15
Solidify Project details for Demo Presentation	2 days	10/16/15	10/19/15
Demo/install Cloudera and Hue	5 days	10/20/15	10/26/15
Gather, collect, and Formatting the data	4 days	10/27/15	10/30/15
Implementation Phase 1 and Graph 1	**20 days**	**10/27/15**	**11/23/15**
Install CentOS Operating System	1 day	10/27/15	10/27/15
Download Hue	3 days	10/27/15	10/29/15
Download and install Cloudera Manager	3 days	10/27/15	10/29/15
Initial Configuration of Cloudera Cluster	6 days	10/30/15	11/6/15
Upload Data in Hue test environment	1 day	11/1/15	11/1/15

Create Hive Database and external tables	7 days	11/1/15	11/7/15
Start configuring Cloudera components	7 days	11/7/15	11/16/15
Develop Queries	2 days	11/7/15	11/9/15
Validate the results	3 days	11/9/15	11/11/15
Visualization of data	4 days	11/11/15	11/14/15
First Demo Presentation	1 day	11/23/15	11/23/15
Implementation Phase 2 and Graph 2	**26 days**	**1/1/16**	**2/5/16**
Add additional nodes	6 days	1/18/16	1/25/16
Establish Visualization software	5 days	1/18/16	1/22/16
Connect Tableau to CDH Environment	5 days	1/25/16	1/29/16
Upload Data in Hue test environment	2 days	1/24/16	1/25/16
Begin configuring YARN/MapReduce 2.0	16 days	1/29/16	2/19/16
Create Hive Database and external tables	2 days	1/24/16	1/25/16
Develop Queries	3 days	1/27/16	1/29/16
Validate the results	2 days	1/29/16	1/31/16
Visualization of data	6 days	1/31/16	2/5/16
Implementation Phase 3 and Graph 3	**17 days**	**2/5/16**	**2/29/16**
Look into other, more innovative visualization software packages	5 days	2/5/16	2/11/16
Format the Data	5 days	2/5/16	2/11/16
Upload Data in Hue test environment	3 days	2/11/16	2/14/16
Create Hive Database and external tables	2 days	2/15/16	2/16/16
Configure other software components in Cloudera	16 days	2/19/16	3/11/16
Develop Queries	3 days	2/17/16	2/19/16

Validate the results	3 days	2/20/16	2/23/16
Visualization of data	4 days	2/24/16	2/29/16
Final Stages	**27 days**	**3/1/16**	**4/6/16**
Testing - Analytics	6 days	3/1/16	3/8/16
Testing - Infrastructure	6 days	3/11/16	3/18/16
Prepare for Demonstration/Final Presentation	14 days	3/18/16	4/6/16

Figure 6, Timeline

Testing Strategy and Plan

During the last phase of our project, testing took place for both team, infrastructure and analytics. For the analytics portion of the project, the main category of testing was data validation. In essence, data validation consists of formatting, importing, querying the data. Our Data Analyst for our project had to download the excel documents with the data, and then properly format the data. As explained above, the formatted data was then imported into Hadoop, which is installed in the Cloudera environment. Then, finally, we tested the queries by using that imported and formatted data. The figure below illustrates the steps taken for each of the data sets used for this project.

Test Results	Pass/Fail	Date Conducted	Able to test in Fall or Spring Semester?	Area tested (Preparing, Ingesting, Querying)
There were a few files that have 2 commas between them. I used Excel to correct this error prior to import	PASS	11/16/2015	Fall Semester	Preparing
Verified that all files were imported	PASS	11/16/2015	Fall Semester	Ingesting
Verfied that data values in both sources were consistent	PASS	11/16/2015	Fall Semester	Ingesting
Verfied that data values in both sources were consistent	PASS	11/18/2015	Fall Semester	Querying

Figure 7, Testing

From the infrastructure side, we needed to ensure that each node was successfully created through the UC Sandbox. This included that each node can successfully connect to the internet, and that each slave node can contact the Master node. We did our network testing by pinging Google and the Master node's IP address. If we got "packet replies" from Google and the Master Node, the node was successful in its network testing. Also, after configuring each slave node, we had to test that each could be successfully added to the Cloudera Manager. To do that, we went through the Cloudera Manager wizard for adding additional nodes. If we went through all the steps in the wizard without any errors, it was added to the Cloudera cluster in the Management console. Once each node was added, we re-ran successful queries to ensure the data was working properly with the newly-added node in the Cloudera cluster. Below are the steps we took for each node created in the Cloudera cluster:

Testing Procedures for Infrastructure
Verify any VM starts in the sandbox with correct OS - CentOS 6.5
Verify Cloudera Manager installs and functions correctly on Master Node (NameNode)
Ping www.google.com from each node to ensure network
Ping the Master Node (10.126.67.210) from each slave node to ensure they are communicating
Verify all services in Cloudera Manager start and stop successfully - there are several dependencies
Run each query from analytics after each node is created to ensure everything still runs smoothly

Figure 8, Testing

Problems Experienced

One problem we have experienced so far is how we are to centrally store the Virtual Machines for our infrastructure. While the sandbox is an option, we feel that performance was not strong enough. So, for now, we are doing everything locally for better performance. However, we do have plans to utilize Amazon Web Services to store all of our data in the cloud.

Also, we had issues of deciding what software packages to apply to the Hadoop File System (HDFS) for importing and analyzing data. Our plan for that has changed a few times, but after finally discussing this with an "expert," we have settled on using Cloudera manager to assist us with the infrastructure development.

And, finally, for this Semester, the analytics need to be created in a non-persistent environment that is only able to be accessed for 3 hours. We will integrate the two environments for the Spring Semester.

Big Data technologies are rapidly evolving. The internet or blog posts from 6 months ago covering Hadoop topics may be out of date (Liekar).

Conclusion

In conclusion, we have illustrated the power of the Hadoop architecture, and defined the process of data ingestion and computation, using analysis of US cities as a use case by setting up a Hadoop cluster and creating 3 analytics: weather, cost of living and job availability.

Bibliography

1. The United States Census Bureau. https://www.census.gov/programs-surveys/acs/about.html
2. CodersVoice Article. http://www.codersvoice.com/a/webbase/install/10/082014/139.html
3. The CentOS forums. https://www.centos.org/forums/viewtopic.php?f=47&t=49428