**分类号**  **学号 I201821251**

**学校代码 密级**

****

**硕士学位论文**

**使用Hadoop生态系统的农业大数据平台**

**案例研究计划和农业经济部-SUDAN**

|  |  |  |
| --- | --- | --- |
| **学位申请人** | **：** | **苏昕** |
| **学科专业** | **：** | **信息与通信工程** |
| **指导教师** | **：** | **刘 澍 讲师** |
| **答辩日期** | **：** |  |

**A Thesis Submitted in Partial Fulfillment of the Requirements for the Degree of Master of Engineering**

**Agricultural Big Data Platform using Hadoop ecosystem**

**case-study Planning and Agricultural Economic Department - SUDAN**

**Candidate : Elsadig Abdullah Osman Adam**

**Major : Information & Communication Engineering**

**Supervisor: Lecturer Shu Liu**

**Huazhong University of Science and Technology**

**Wuhan 430074, P.R.China**

**May, 2020**

**独创性声明**

本人声明所呈交的学位论文是我个人在导师指导下进行的研究工作及取得的研究成果。尽我所知，除文中已经标明引用的内容外，本论文不包含任何其他个人或集体已经发表或撰写过的研究成果。对本文的研究做出贡献的个人和集体，均已在文中以明确方式标明。本人完全意识到本声明的法律结果由本人承担。

学位论文作者签名：

日期：年月日

**学位论文版权使用授权书**

本学位论文作者完全了解学校有关保留、使用学位论文的规定，即：学校有权保留并向国家有关部门或机构送交论文的复印件和电子版，允许论文被查阅和借阅。本人授权华中科技大学可以将本学位论文的全部或部分内容编入有关数据库进行检索，可以采用影印、缩印或扫描等复制手段保存和汇编本学位论文。

保密□，在年解密后适用本授权书。

本论文属于

不保密□。

（请在以上方框内打“√”）

学位论文作者签名： 指导教师签名：

日期：年月日 日期：年月日

# 摘要

**关键词**：大数据集成、大数据存储、大数据分析、农业

# Abstract

**Keywords**: Big Data Integration, Big Data Storage, Big Data Analysis, Agriculture

**Contents Table**

[摘要 I](#_Toc37237299)

[Abstract II](#_Toc37237300)

[Figures List V](#_Toc37237301)

[Tables List VIII](#_Toc37237302)

[1 Introduction 1](#_Toc37237303)

[1.1 Thesis Motivation 2](#_Toc37237304)

[1.2 Literature Review 2](#_Toc37237305)

[1.3 Problem Statement 4](#_Toc37237306)

[1.4 Research Question 5](#_Toc37237307)

[1.5 Thesis Objectives 5](#_Toc37237308)

[1.6 Research Goal 6](#_Toc37237309)

[1.7 Research Methodology 7](#_Toc37237310)

[1.8 Proposed System 8](#_Toc37237311)

[1.8.1 Data Ingestion 8](#_Toc37237312)

[1.8.2 Data Storage 9](#_Toc37237313)

[1.8.3 Analyze and visualization 10](#_Toc37237314)

[1.9 Thesis Organization 10](#_Toc37237315)

[2 Background 12](#_Toc37237316)

[2.1 The Planning and Agriculture Economic Department (PAE) 12](#_Toc37237317)

[2.1.1 Data Sources Partners 12](#_Toc37237318)

[2.1.3 Data Integration and Dependencies 14](#_Toc37237319)

[2.2 Big Data 15](#_Toc37237320)

[2.3 Hadoop 15](#_Toc37237321)

[2.3.1 Hadoop Core Components 17](#_Toc37237322)

[2.3.2 Hadoop Ecosystem 20](#_Toc37237323)

[2.3.3 Cloudera Distributed Hadoop (CDH) 26](#_Toc37237324)

[2.4 Hadoop for agriculture 27](#_Toc37237325)

[2.4.1 The big data benefit in Agriculture sector 27](#_Toc37237326)

[2.4.2 Big Data in Agriculture Use Case 28](#_Toc37237327)

[2.4.3 What is Big data can offer for agriculture: 29](#_Toc37237328)

[3 Data Recognition 31](#_Toc37237329)

[3.1 The current system in the Planning and Agricultural Economics 31](#_Toc37237330)

[3.1.1 The Text files 31](#_Toc37237331)

[3.1.2 The MySQL Database 32](#_Toc37237332)

[3.1.3 The weather Data 33](#_Toc37237333)

[3.2 Data structure 33](#_Toc37237334)

[3.2.1 CSV data structure 33](#_Toc37237335)

[3.2.2 The relational database data structure 36](#_Toc37237336)

[3.2.3 Data structure of Weather API 38](#_Toc37237337)

[3.3 Data Sources 41](#_Toc37237338)

[3 System Design and Data Structure. 44](#_Toc37237339)

[3.1 Data Structure Design 44](#_Toc37237340)

[4.1.1 Data ingestion 44](#_Toc37237341)

[4.1.2. Data storage 45](#_Toc37237342)

[4.1.3 Analysis and visualization 46](#_Toc37237343)

[4.2 Development Tools 47](#_Toc37237344)

[4.2.1 Development tools for Data Ingestion Layer 47](#_Toc37237345)

[4.2.2 Development tools for the Data Storage layer 48](#_Toc37237346)

[4.2.3 Development tool for the Data Analysis and Visualization layer 48](#_Toc37237347)

[4.2 System implementation 50](#_Toc37237348)

[4.2.1 Implementing the base machine 50](#_Toc37237349)

[5. System Implementation 65](#_Toc37237350)

[5.1 Introduction 65](#_Toc37237351)

[5.2 Common Family Planning Methods in Sierra Leone 66](#_Toc37237352)

[5.3 Relationship on family planning methods among the three data sources. 67](#_Toc37237353)

[5.4 Data Selection 68](#_Toc37237354)

[5.5 Analysis and Visualization 73](#_Toc37237355)

[5.5.1 Goal of the analysis 73](#_Toc37237356)

[5.5.2 Analysis to determine the level of inconsistency of family planning data 76](#_Toc37237357)

[5.5.3 Investigating possible factors leading to the inconsistency of family planning data 80](#_Toc37237358)

[5.6 Challenges 95](#_Toc37237359)

[5.7 Conclusion 97](#_Toc37237360)

[Acknowledgement 99](#_Toc37237361)

[REFERENCES 100](#_Toc37237362)

[Appendix Abbreviations List 105](#_Toc37237363)

# Figures List

[Figure1. 1 Distribution of Health Commodities 5](#_Toc10542642)

[Figure1. 2 Proposed system development phases 11](#_Toc10542643)

[Figure2. 1 Reporting flow in the DHIS2 14](#_Toc10542644)

[Figure2. 2 Reporting flow in the IHRIS 15](#_Toc10542645)

[Figure2. 3 Reporting flow in the LMIS 16](#_Toc10542646)

[Figure2. 4 a hierarchical structure of organization units 17](#_Toc10542647)

[Figure2. 5 Data elements with category options and data values 18](#_Toc10542648)

[Figure2. 6 Entity Relationship Diagram of the DHIS2 19](#_Toc10542649)

[Figure2. 7 Entity Relationship diagram of the IHRIS. 22](#_Toc10542650)

[Figure2. 8 Data structure of the LMIS 24](#_Toc10542651)

[Figure2. 9 Common attributes among the three systems 26](#_Toc10542652)

[Figure3. 1 Data Integration Subsystem 29](#_Toc10542653)

[Figure3. 2 Data Storage Subsystem 30](#_Toc10542654)

[Figure3. 3 Data Analysis and Visualization Subsystem 32](#_Toc10542655)

[Figure3. 4 General structure of the proposed platform 33](#_Toc10542656)

[Figure3. 5 process group with processors 36](#_Toc10542657)

[Figure3. 6 relationship between processors in Apache Nifi 36](#_Toc10542658)

[Figure 4. 1 A window showing Apache Nifi successfully started 43](#_Toc10542659)

[Figure 4. 2 Database controller service to connect to the DHIS2 data source. 44](#_Toc10542660)

[Figure 4. 3 a QueryDatabaseTable processor to extract records from the organisationunit table in the external database and ingest to Apache Nifi. 45](#_Toc10542661)

[Figure 4. 4 PutHiveStreaming processor to ingest data to Hive 47](#_Toc10542662)

[Figure 4. 5 Relationship between the QueryDatabaseTable processor and the PutHiveStreaming processor 48](#_Toc10542663)

[Figure 4. 6 flow with processors to ingest data to the main repository from the DHIS2 external data source**.** 49](#_Toc10542664)

[4. 7 flow with processors to ingest data to the main repository from the DHIS2 external data source**.** 50](#_Toc10542665)

[Figure 4. 8 Controller service to connect to the IHRIS external data source. 51](#_Toc10542666)

[Figure 4. 9 QueryDatabaseTable processor to ingest data to Apache Nifi form the hippo\_facility\_type table in the IHRIS data source. 52](#_Toc10542667)

[Figure 4. 10 PutHiveStreaming processor to stream data IHRIS data to Hive 53](#_Toc10542668)

[Figure 4. 11 flow with processors to ingest data to the main repository from the IHRIS external data source 54](#_Toc10542669)

[Figure 4. 12 flow with processors to ingest data to the main repository from the IHRIS external data source 55](#_Toc10542670)

[Figure 4. 13 GetFile processor to get LMIS data to Apache Nifi 56](#_Toc10542671)

[Figure 4. 14 InferAvroSchema processor to examine the content of the file 57](#_Toc10542672)

[Figure 4. 15 ConvertCSVToAvro processor to convert the CSV file to AVRO format 58](#_Toc10542673)

[Figure 4. 16 PutHiveStreaming processor to ingest LMIS data to HIVE 59](#_Toc10542674)

[Figure 4. 17 a complete flow with processors to ingest LMIS data to the main repository 60](#_Toc10542675)

[Figure 4. 18 Data ingested to Apache Hive 62](#_Toc10542676)

[Figure 5. 1 A flow chart illustrating conditions to determine whether the data is consistent or not 74](#_Toc10542677)

[Figure 5. 2 a comparison between family planning cases and consumption of family planning health commodities for Pujehun district. 76](#_Toc10542678)

[Figure 5. 3 a comparison between family planning cases and consumption of family planning health commodities for Tonkolili district. 78](#_Toc10542679)

[Figure 5. 4 a comparison between the oral contraceptive cases, consumption, difference and cases whose consumption are not reported for Pujehun district per month. 82](#_Toc10542680)

[Figure 5. 5 comparison between the quantity of oral contraceptive pills in store and the quantity dispensed. 83](#_Toc10542681)

[Figure 5. 6 a comparison between implant cases, consumption, difference and number of implant cases whose consumption are not reported for Pujehun district per month. 86](#_Toc10542682)

[Figure 5. 7 comparison between the quantity of Jadelle implants in store and the quantity dispensed. 86](#_Toc10542683)

[Figure 5. 8 comparison between injectable cases, consumption, difference, and number of injectable cases whose consumption are not reported for Pujehun district. 87](#_Toc10542684)

[Figure 5. 9 comparison between the quantity of Injectable (i.e., Depo Provera) in store and the quantity dispensed. 88](#_Toc10542685)

[Figure 5. 10 comparison between cases, consumption, difference, and number of cases whose consumption are not reported for male condoms in Pujehun district. 89](#_Toc10542686)

[Figure 5. 11 comparison between the number of male condoms in store and the quantity dispensed. 90](#_Toc10542687)

[Figure 5. 12 comparison between cases, consumption, difference, and number of cases whose consumption are not reported for female condoms in Pujehun district. 91](#_Toc10542688)

[Figure 5. 13 comparison between the number of male condoms in store and the quantity dispensed. 92](#_Toc10542689)

[Figure 5. 14 comparison between cases, consumption, difference, and number of cases whose consumption are not reported for Injectable in Pujehun district. 93](#_Toc10542690)

[Figure 5. 15 comparison between the quantity of Injectable in store and the quantity dispensed. 94](#_Toc10542691)

# Tables List

[Table1. 1 Work plan for the development of the proposed platform 12](#_Toc10542692)

[Table3. 1 Apache Nifi concepts 35](#_Toc10542693)

[Table 4. 1 Configurations on the jdbc interpreter to connect to the HIVE metastore 63](#_Toc10542694)

[Table 4. 2 Configurations on the Spark interpreter to connect to the HIVE metastore 63](#_Toc10542695)

[Table 5. 1 Family planning services and health commodities 73](#_Toc10542696)

[Table 5. 2 Cases, consumption, difference and percentage difference of family planning data for Pujehun district. 76](#_Toc10542697)

[Table 5. 3 Cases, consumption, difference and percentage difference of family planning data for Tonkolili district. 77](#_Toc10542698)

[Table 5. 4 cases, consumption, and difference for Oral Contraceptive pills in Pujehun district 81](#_Toc10542699)

[Table 5. 5 number of oral contraceptive pills cases whose consumption are not reported in the LMIS and its percentage for Pujehun district. 82](#_Toc10542700)

[Table 5. 6 cases, consumption, and difference for Implants family planning service in Pujehun district 84](#_Toc10542701)

[Table 5. 7 number of implant cases whose consumption are not reported in the LMIS and its percentage for Pujehun district. 84](#_Toc10542702)

[Table 5. 8 cases, consumption, difference, number of cases for injectable whose consumption are not reported in the LMIS and its percentage in Pujehun district. 87](#_Toc10542703)

[Table 5. 9 cases, consumption, difference, and cases whose consumption are not reported in the LMIS and its percentage for male Condoms in Pujehun district. 89](#_Toc10542704)

[Table 5. 10 cases, consumption, difference, and cases whose consumption are not reported in the LMIS and its percentage for female Condoms in Pujehun district. 91](#_Toc10542705)

[Table 5. 11 cases, consumption, difference, cases whose consumption are not reported in the LMIS and its percentage for Injectable in Pujehun district. 93](#_Toc10542706)

# 1 Introduction

The agricultural sector is the core of Sudanese population and the primary source of livelihoods for a majority of its citizens. The Sudanese economy is predominately rural with 65.1% percent of the population deriving their livelihoods in rural areas. The agricultural sector contributes more than one third to GDP (39.6%) and more than two third (80%) to labor force (CIA, 2019).

However, Sudan like many other African countries seems to be afflicted with persistent food fluctuations over the last two decades.

It is sometimes alleged that prices of food crops in the Upper East Region are unduly depressed in the post-harvest period and that they rise to excessive heights in the period just before harvest. This large increase in prices is attributed to heavy storage losses, exploitative speculation and simple improvidence. To the extent that it is true, it may reduce farmers’ incomes and thus their incentive to produce and even provoke actual food shortage. Severe hunger may result when farmers, who have sold their crops in the glutted post-harvest markets, find it necessary to buy them back at two, three, or four times the price in order to feed their families while waiting for the new crops to mature. The need for a scientific research that will provide evidence to show the direction of price movement cannot be overemphasized.

The study seeks to use sales (prices of food crops) data as a case study to compare the performance of these two methods based on the values of their Mean Absolute Percentage Error (MAPE), Mean Squared Deviation (MSD) and Mean Absolute Deviation (MAD) and to confirm or otherwise the validity of the statement that Holt-Winters model performs better when both seasonality and trend are present. The study seeks to examine the trend of the prices of the food crops under study within the study period (2008-2018) and try to forecast the prices for the next twelve months. It must be emphasized that one needs not expect accurate forecast in this kind of exercise. Even with the original data, there have been serious outliers, which only distort the trend.

Although the researcher has chosen prices of three cereals crops as a case study, the prediction of prices of crops are by no means can apply on other crops. It is our hope that the exposition of this problem and perhaps its effects will serve as a basis for further research so that more comprehensive reports and recommendations can be made to the Economic Planners of Sudan to prevent the occurrence of such situations in the future.

The study reviews the prices of some of the cereals crops in the Sudan. Primarily the study is to afford any planner the opportunity of knowing what to expect of the prices of the cereals crops and thus determine methods of adjusting to the situation prevailing if the factors remain the same.

The main aim of this paper is to build integrated platform to ingest, store, and visualize the data beside construct a statistical model for analyzing and forecasting prices of cereals crops in Sudan. In doing this the study will examine the trends in the prices of cereal crops and their prospects for future, it also seeks to build an exponential smoothing model to analyze the prices of food crops in Sudan and to compare the efficiency of Holt-Winters Multiplicative and Double Exponential Smoothing models in forecasting prices of cereal crops.

The main aim of this paper is to build integrated platform to ingest, store, and visualize the data beside construct a statistical model for analyzing and forecasting prices of cereals crops in Sudan. In doing this the study will examine the trends in the prices of cereal crops and their prospects for future, it also seeks to build an exponential smoothing model to analyze the prices of food crops in Sudan and to compare the efficiency of Holt-Winters Multiplicative and Double Exponential Smoothing models in forecasting prices of cereal crops.

## 1.1 Thesis Motivation

Collecting and organizing data for use in decision-making through foresight, analysis and prediction is a major challenge. We face this problem periodically in organizations that use a huge amount of data. With the opportunity provided by big data to ingest, store, analyze and visualize data, we believe that building an integrated platform will be a great help to get the most out of the data.

## 1.2 Literature Review

Van-Quyet Nguyen etal[[[1]](#endnote-1)] Focusing on designing a platform for collecting and analyzing agricultural data they use flume to collect a real time data from sensors and web pages, and they use HDFS for storing data after clean it beside hbase as data model and hive as data warehouse to process structured data. For static data they use Sqoop to import data from CSV, and XLS files and to export the result into MySQL. For analyzing they use spark after clustering and classifying with K-Means and Naïve Bayes. We will use different tools with limit feature. We will use HDFS for storing and NiFi for collecting Static and Real time data. we will focus on collecting static data where our real time data will be the metrology data only.

When Barghash [[[2]](#endnote-2)] try to build interactive platform for agricultural data in Egypt, he build a relational schema to host subset of agri-data and analyze it then he test it with some queries and make visualization for that to show the penefit of storing and analyzing data from farm plus the wither factor. He use Hive to connect to ODBC sources and PowerBI for visualization. We can consider his model as traditional ETL and he depend on Historical metrology data for storing he use HDFS with hive on other hand we will build a data hub as we explain in the different between ETL and data hub.

In Twards a big Health data analytics platform [[[3]](#endnote-3)] they propsed data analysis pipeline to selecting, integrating, analyzing, visualizing, and sharing the healthcare data. They have very good platform doing standardization using SNOMED-CT as standard data form, their primary focus on data standardization to integrate data from multiple sources. For that they use MapReduce with HDFS their platform offer standardization and integration to prepare the integrated dataset, and data selection function with Hive query to analysis data which done with R, Impala and RImpala, also they offer web base data visualization with Shiny.

Because of the lack of sources that talk about big data in agriculture, we have benefited from a paper[[[4]](#endnote-4)] related to the health sector as well He provided insight into famous initiatives in big data for healthcare and steps to processs the analysis, the benefits and the challenge.

On big scale also there are agINFRA and CGIAR. agINFRA is a project funded by European Commission aim to introduce agricultural scientific communities to the vision of open and participatory data-intinsive science [[[5]](#endnote-5)] and they use dedicated infrastructures.

We use Megan Stubbs paper [[[6]](#endnote-6)] as a guide to classify our parceners to a category.

At the beginnig we had have a Data warehouse and we organize data on it using ETL. But that still very painful and lead to fail 50% of all data[[[7]](#endnote-7)]. For that a new technology introduced to solve this headache like virtual database, data lake and data hubs. Where Virtual database also called a Federated Database is just a system that accept a query and invented to be a big data. Data lake appear with Hadoop it is when you move your data from all sources to one system (HDFS) without any need to normalizing, indexing, or structuring the data. In the other hand in data hub we need to move data physically and indexed it, and unlike data like it would support discovery, indexing and analytics\cite{marlogic}. In what we can call Movement, Harmonization, and Indexing operations.

Enterprise face many challenges in data integration effort [[[8]](#endnote-8)] they have large amount of data, with different technologies in multiple formats within many data schema and rapidly changing in quantity and quality. We can say the limiting nature of traditional technologies renders it ineffective as an enterprise integration pattern in the modern era.

Kemper and etal [[[9]](#endnote-9)] a great insight about the benefit of moving from warehouse to data hub. Hadoop introduce Enterprise data hub (EDH) as central repository for structured, semi-structured, and unstructured data.

## 1.3 Problem Statement

The Department of Planning and Agricultural Economics (PAE) faces a monthly dilemma in preparing an integrated report on agricultural operations. The problem comes from the fact that the data required to complete this report are spread over more than fifteen different bodies, and of course, each is located in a separate location. In addition, the Department of Planning and Agricultural Economic use an Excel files only to provide this data, although the Department is now seeking to create a database for that, but still they will face same problems. All this affects the integrity and the analysis output of these data.

PAE needs to collect data from the various departments of the Ministry of Agriculture on the crops amounts, places of cultivation, seeds sufficient for the agricultural season, whether there are sufficient quantities of pesticides, the amount of feed expected to be consumed or produced, and the needs of each sector (whether rainy, irrigated, or horticultural sector). The Ministry of Animal Resources and Fisheries provide the number of herd and quantities of animal production. The Ministries of Interior and Bank of Sudan, prepare the volume of internal and intra-trade for both agricultural and animal products with financing loans. the General Authority for Meteorology collects weather statistics in those areas. In addition to other data they need to make recommendations to decision makers or to reflect the full picture in front of them.

The absence of a unified platform for data collection, ingestion, storage and visualize in scientific and acceptable form to high management and stakeholders is a chronic problem faced by management and related entities.

## 1.4 Research Question

In this study we want to answer the question:

1. How can we build a platform that helps us to integrate, ingest, store, visualize, and analysis data, for the agricultural sector (Sudan).
2. How can we visualize the data for some important cereals. so that the Department and all other relevant bodies benefit in the process of obtaining and analyzing information later.

## 1.5 Thesis Objectives

By ingesting, filtering, cleaning, harmonizing, storing, and visualizing the agricultural data integrated with factor linked to production and consumption, we give a lot of stakeholder, analysts and decision maker ability to get benefit from the data. Because we will provide them a single platform where there can found all the relevant data that they want. To make a smart decision, make a precise analysis, or save a money and resources.\\

This proposed platform contains a number of subsystems necessary to obtain a highly efficient and useful system. These subsystems contain a subsystem for: data ingestion, data cleaning, data storage, data analysis, predicting and finally data visualization.\\

* **General Objectives**

1. Building Data Hub
2. Analyzing the production changes of three cereal crops comparing to weather
3. Design Data visualization Front end

* **Specific objective**

1. Build an integrated platform that collects data from different system data sources at regular intervals (CSV, DB, PDF, API).
2. Deploy and configure a platform to communicate with the repository and ensure that the integrated data is selected, analyzed and visualized.
3. Deploy a data hub platform that acts as a repository for storing data extracted from different systems.
4. Comparing the change in production for three cereal crops (Sorghum, Millet and Wheat) using hive under Hue GUI.
5. Analyze and provide visualization.

## 1.6 Research Goal

The aim of the research is to maximizes the contribution of the Department of Planning and Agricultural Economics in development and planning by:

1. provide a cooperative platform that collects data from different sources in different format
2. Clean and store the data in one place to allow specialists in the agricultural field and decision makers to get integrated data. that enabling them to carry out analysis, required to reach accurate forecasts and correct decisions.
3. Provides visualization of data that allows non-specialists to understand the general direction of data and get out of it with quick decisions.
4. Analysis the relation between precipitation amount of rain and production for four crops.

We can say that our work is focused on flow management rather than Stream management. In **Flow Management**, we will do the following:

* **Acquire**: keeps track of data from several sources and several types
* **Move and filter**: Sometimes we may need to send data from client X to be a input to another instance.
* **Intelligent routing**: Routing and filtering rules will change data as it moves at run time
* **Deliver**: data deliver to more than one downstream user.
* **Parse** the conversion of data to a standard format that downstream system can understand
* **Enrich** the data by adding contextual information such as weather conditions etc.

What we will do on the side of the **Manage stream** will be limited to three points:

* **Join and split** data from more than one source.
* **Score/execute** some analytical model on individual event.
* **Custom dashboard/graph** to analyze time series data and insights / alerts.

## 1.7 Research Methodology

We took five steps to build the platform that we show in this paper:

* Data Discovery
* Architectural Design
* Implementation
* System Testing
* Data Analysis and Visualization

In the **exploration of data**, we will look at the systems in which affiliates provide their data and look at these data to discover patterns.

In the **structural design** step, we will look at the structure of the proposed system by looking to proposed subsystems (collection, clean, predicting, analysis and visualizing) and tools that we will use and why we chose these tools.

In **implementation** we will setting up of the infrastructure to develop the proposed platform. we will focus on requirement and how to configure the devices and step to install required system and tools. in addition to building process to ingest and storage the data from different sources to the repository. and we will provide the steps to build the dashboard to visualize the data.The **testing** step assesses the software for errors and documents bugs if there are any, so we can ensure that our platform work properly.

at the last we reach the **Analysis** **and** **Visualization** step. where we can analysis and visualization some of agricultural factors. in addition, we will provide algorithm to predict price for main three cereal crops in Sudan. to provide food security department with most accurate forecast for those cereals. We will use time series (holt-Winters) comparing with simple linear regression. And we will present forecasting for rain fall also using time series. We will show all these result and information using visualization to make it easier to understand.

## 1.8 Proposed System

We propose an integrated system to ingest, store, analyze, and visualize data. We can discribe it as it illustrated in figure 1.1.

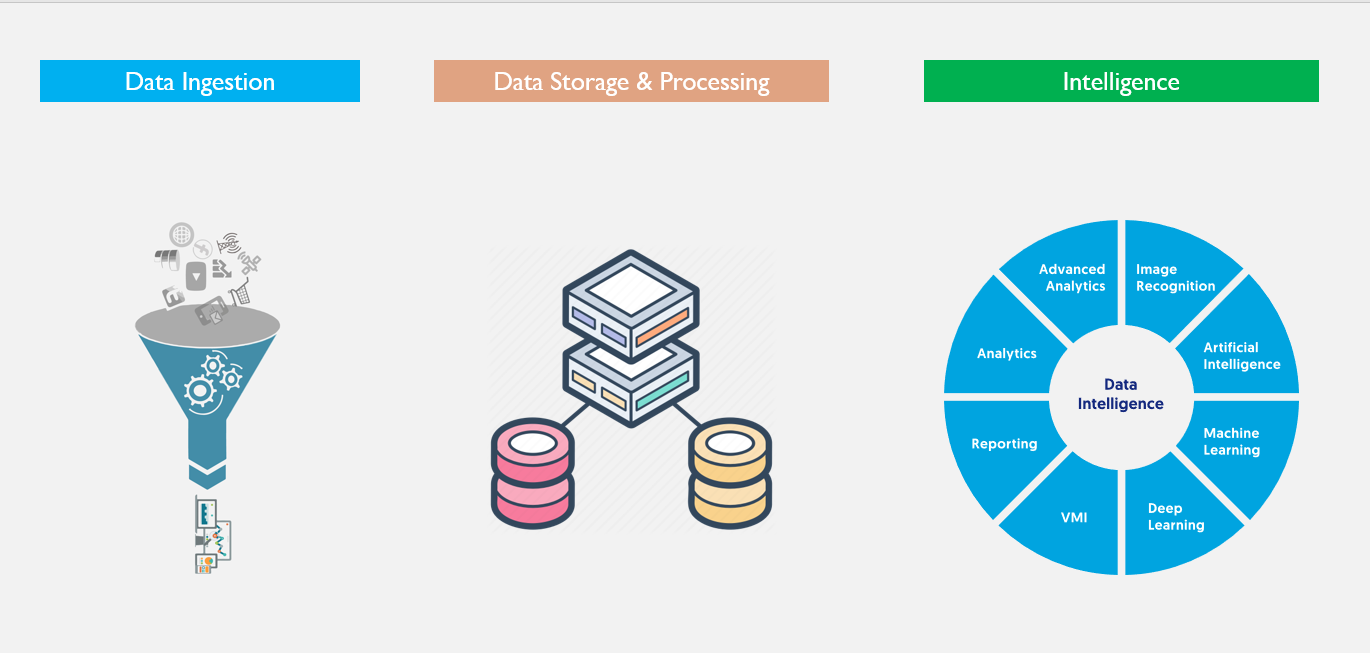


Figure 1.1 the structure of proposed system

1.8.1 Data Ingestion

In ingestion subsystem we will use a different tool from Hadoop ecosystem, those tools include: Sqoop, NiFi, Kafka as we show in figure 1.2

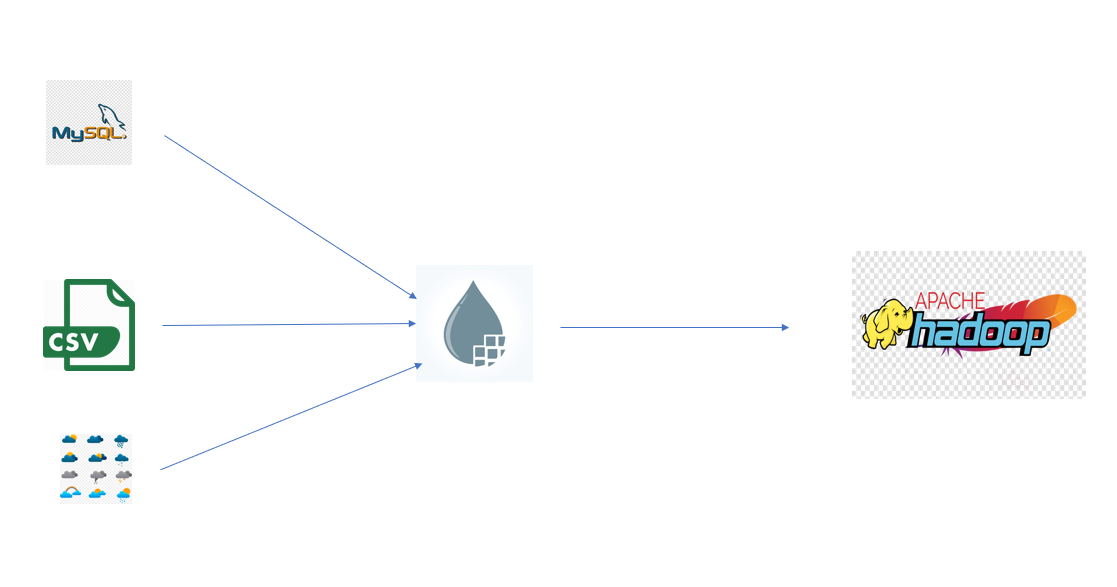


Figure 1.2 the ingestion subsystem components

1.8.2 Data Storage

For data storage we use Hadoop HDFS for storing, with hive for relational data and also we will use MapReduce to make transforming operation on data if needed as in figure1.3.



Figure 1.3 Storage subsystem components

1.8.3 Analyze and visualization

In this stage we will use Hue with Apache Hive for analysis and visualization, we will use Apache Spark under Ambari with evolving oozie also.



Figure 1.4 Analysis and visualize subsystem

## 1.9 Thesis Organization

We have divided this study into seven chapters as follows:\\

**Chapter 1: introduction** where we provided a background on the big data and the platform with the research questions, research problems, research objectives and research methodology.

**Chapter 2: Background** where we discussed some of the previous research in the field of big data platforms. And about the benefit of the agricultural field from the big data. Beside research on Data analysis and visualization. The use of time series on agricultural prediction especially Holt-Winters method.

**Chapter 3: Data Recognition** We studied the data derived from the current system used in the Department of Planning and Agricultural Economics to determine the structure of the data and the relationships between them. With the benefit of integrating them into a single platform.

**Chapter 4: Design and Data Structur** in the design section we present the proposed platform and the tools we have used to accomplish this.

**Chapter 5 implementation** The researcher discussed the method of implementing subsystems to ingest, clean and store data to be ready for analysis and presentation.

**Chapter 6: Experiments and Result** In this chapter we used the platform we have built, to ingest, store, analyze, and visualize the data available on the platform with a case study on production for three main cereals in Sudan.

For conclusion, we discussed the evaluation results obtained from the platform tested with the case study.

# 2 Background

## 2.1 The Planning and Agriculture Economic Department (PAE)

The Department of Planning and Agricultural Economics communicates with more than a dozen departments or ministries to be able to provide a good report to help in decision-making, And to carry out scientific planning based on real survey data and statistical real.

2.1.1 Data Sources Partners  
The reports submitted by the Department of Planning and Agricultural Economics are important because they contain related information that is of interest to these entities separately. These  
Entities include  
1. Agricultural Economics Department: It is affiliated to the Ministry of Agriculture  
and provides production costs in addition to costs and marketing margins prices of agricultural crops, resource economics feasibility studies and competitiveness and relative Studies.  
2. Department of Agricultural Statistics: It reports to the Ministry of Agriculture and provides data on cultivated and harvested areas, and production and productivity statistics for the main crops in the country, such as Sorghum, Millet, Wheat, Maize, Rice, Sesame, Peanuts, Sunflower, Cotton, Hibiscus, and Melon seeds.  
3. Department of Food Security: the fourth body of the Ministry of Agriculture and  
Forestry. They provide state-level crop statistics and annual food security indicators.  
4. Gender Integration Unit: This unit is also affiliated to the Ministry of Agriculture  
and Forestry and provides the participation rate of women in agriculture and projects to Increase the income of rural women.  
5. Pasture and Fodder Department: It is Department in the Ministry of Agriculture and Forestry provides data on natural and cultivated rangelands as well as estimates of Fodder consumption and estimates of plant waste as fodder.  
6. Department of Seeds: This department belongs to the Ministry of Agriculture and Forestry, which provides an improved and approved seed quantity.  
7. Plant Protection Directorate: The department provides statistics on pesticide inspection and pest control products. In addition to what is allowed to be imported from pesticides.  
8. The National Forestry Commission (NFA): It is an agency of the Ministry of Agriculture and Forestry that tracks reserved and unlisted forest areas in addition to assessing forest production and types.  
9. Management of the horticultural sector: there provide the PAE with statistics for a Horticultural sector production.  
10. Irrigated Sector Management: It assesses the needs of the irrigated sector as well as surveys and production plans for the sector.  
11. Management of the rain sector: It assesses the rain sector needs, as well as statistics and productivity plans.  
12. Central Bureau of Statistics: Contributes to providing population, agricultural census, gross domestic product, inflation rates, and household budget surveys.  
13. General Directorate of Customs Police: Provides statistics of foreign trade and Intra trade.  
14. Bank of Sudan: We obtain statistics of internal and external trade in addition to loans and agricultural financing.  
15. Ministry of Livestock and Fisheries: From this ministry, we can find estimates of the national herd, estimates of milk and meat production, slaughter numbers, and estimates of poultry production.  
16. Ministry of Irrigation: The most important statistics received from the ministry of Irrigation is the annual supplemental seasonal, permanent and groundwater in addition to the annual consumption of water in agriculture and estimating the water needs of the Irrigated sector.  
17. General Meteorological Authority: from which we get weather forecasts and monitoring in the country.  
2.1.2 Data Flow  
All of these entities participate in providing data to the PAE Department and at the same time represent the largest data consumer in the platform



Figure 2.1 *The Data source for the Department of Planning and Agricultural Economics*

Making this data distributed in departments and ministries, integrated leads to optimization. However, we can say that leaving each set of data individually is a clear violation of data integrity. While each section needs data from the other. Waiting for this data to be collected, analyzed, and explored may cause some deficiencies.

2.1.3 Data Integration and Dependencies

For example, department of pasture and forage in order to estimate the amount of feed expected to be consumed, they should have the number of herd estimates from the Ministry of Livestock and Fisheries. The Ministry of Livestock and Fisheries needs to know whether the available feed is sufficient to feed the herd or if it needs to request additional feed.

The Customs Authority needs to know the types of seeds allowed to enter the country and whether the importer has a record in the department of seed. In contrast, the department of seed needs to know the amount of seeds imported.

We find that the relationship between these entities is symbiotic relationship each side needs information from the others.

## 2.2 Big Data

Nowadays, the data growing very fast the amount of data generated exceed 30,000 GB every second from various sources. We have pushed the traditional database system to the limit and face failure to scale. So we found ourselves in need of new technology and here came big data. The pioneer company in the field were Google presenting distributed filesystem, MapReduce, and distributed locking services. Amazon with Dynamo, which is a distributed Key/value store. Then the open source community provide us with Hadoop, HBase, MongoDB, Cassandra, and Countless projects [[[10]](#endnote-10)]

Big data is the description of a large amount of either organized or unorganized data that is analyzed  
To make an informed decision or evaluation. explain big data adequately, there are three main characteristics related to the big data: volume, Variety, and velocity

**Volume**: use to describe the enormous amount of data that used by big data, and to name a system as a big data system it should be able to handle this amount of data and its expected growth

**Variety**: the types of sources used by analytics big data systems with different storage formats wither it is structured or unstructured format.

**Velocity**: describe how fast the data had been generated. And the speed to processed this generated data. To ensure that data is current and updated in real-time[[[11]](#endnote-11)].

And we can add **veracity** as forth characteristic, which is the quality of data by cleaning the data to improve the accuracy. [[[12]](#endnote-12)]

By 2017 there are roughly 100 projects related to big data or Hadoop. Apache produces a lot of open sources solution to solve a new challenge or to solve an old challenge in a new way. They introduce Hive as an extensive database, Kafka for messaging, and more [[[13]](#endnote-13)]

As Forbes mention, we generate almost 2.5 quintillion byte of data every day, and it expected to increase dramatically with the rise of IoT. With all this amount of data, we have to ask our self, how we can make sense of these massive amounts of data, analyze, and get actionable intelligence out of it.

## 2.3 Hadoop

As a solution from Yahoo to improve its search platform, they introduce Hadoop in 2006, which become a defacto standard in the industry, helping a lot of enterprises to get benefit from their big data. When we say Hadoop, now we don't mean just core apache Hadoop project but also apache Hadoop technology along with an ecosystem that works with.

According to[[[14]](#endnote-14)], we can define Hadoop as "Hadoop is a platform that provides both distributed storage and computational capabilities”. The figure 1.1 shows the main components of Hadoop

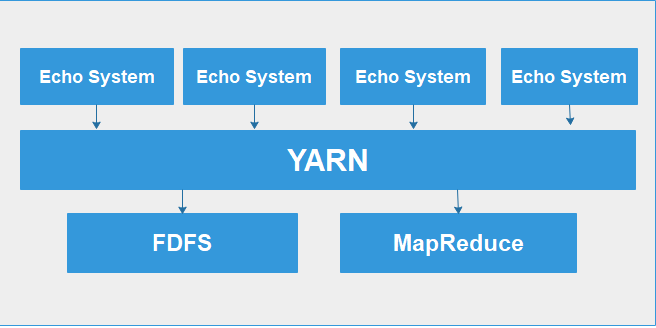


Figure 2.1 the high level for Hadoop components

Hadoop came with the term Schema-on-read which mean you would load your data first then ask question later in contrast for schema-on-write in a relational database when the schema encoded when the data stored in analytic platform.

Benefit of Hadoop

1. **Accessibility**: it runs on large clusters either on a local machine or on cloud
2. **Robust**: because its aim to run on a different platform
3. **Scalable**: we can add more than node to the cluster
4. **Simple**: easy to use where user can writ efficiently parallel code quickly

in this research, we will use the following systems/tools

* 1. HDFS
  2. Hive
  3. NiFi
  4. Spark

**The component of Hadoop**

Hadoop is made up of two main components, So at the high level, we can divide Hadoop to tow main components:

2.3.1 Hadoop Core Components  
it is the platform to run and execute distributed computing tasks and content of three major components:

* + - 1. **HDFS (Hadoop Distributed File System**): a distributed filesystem. Stores data in a distributed, scalable, fault-tolerant manner. Implemented as master-slave architecture where NameNode is the master and DataNode (one or more) are the slaves. The figure 1.2 show digram for HDFS structure. The dataNode program manages each data node. In the name node, we found information about where to store and retrieve data besides the number of copies of the data. By default, HDFS stores three copies of the data across the cluster. It's modeled after Google File System (GFS) paper. The key traits for it are scalability and availability, with data replicant to achieve fault tolerance.

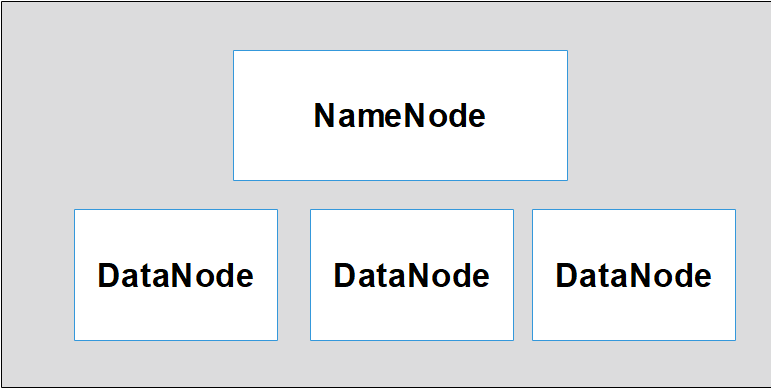


Figure 2.1 HDF structure of NameNode and DataNode

* + - 1. **YARN (Yet Another Resource Negotiator**): is a distributed resource scheduler to keep track of CPU, RAM, and Disk spaces, in order to smooth the execution of tasks. It introduced with Hadoop 2. It enables us to execute a new set of workloads as natively supported.  
         it appears as apart of Hadoop in 2012, it is main rule is resource negotiation which before depends on the operating system for each node on the cluster. YARN enables us to execute a new set of workloads as natively supported. Its contents of a client, which uses a resource manager that connects with Node manager and application master this gives us to allocate containers in various configurations.
      2. **MapReduce**: A distributed data processing and analysis framework. To analyze data stored in HDFS. It consists of tow processing phase: map and reduces. In the map phase, it performs record by record; in a reduce phase, it makes aggregation operation on a group of records. MapReduce gets the data from the map to reduce functions by shuffle and sort. If you find MapReduce too abstracted you can use Crunch, Cascading, or Scalding tools and also there is a high-level language like Apache Hive or Apache Big. MapReduce is a batch-based framework to distributed computing and also after Google's paper on MapReduce. With MapReduce, we can parallelize work over a numerous amount of row data. Figure 2.3 show the process flow for MapReduce.

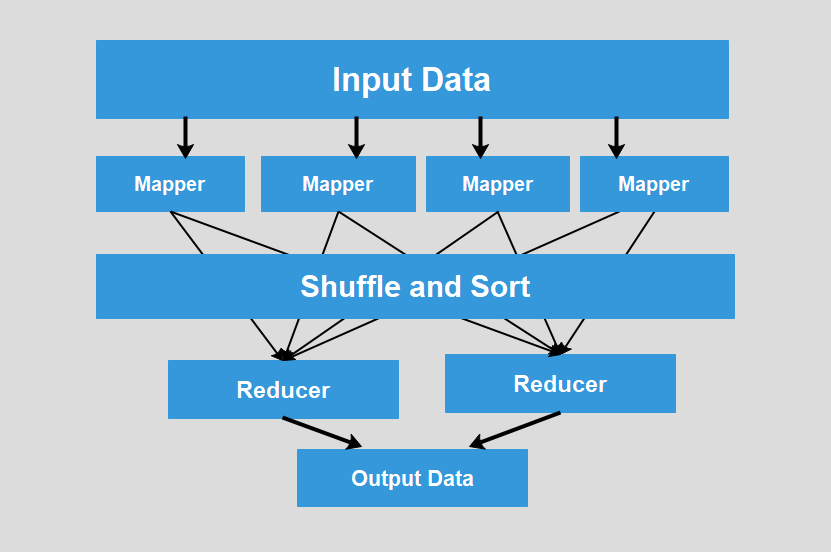


Figure 3.3 the process flow for MapReduce

* + - 1. **Cluster**

The building blocks of Hadoop is cluster when we configure cluster which running a set of daemons (resident programs) on the different servers. The daemons have a rule; some on a server only and others may exist across multiple servers, the daemons like:

* **NameNode** (NN): it directs the slave DataNode daemons to perform the low-level I/O tasks. It keeps track of the file blocks in DataNode. Each cluster had one NN
* **DataNode**: it is the actual files, whereas reading and writing the file we distributed on HDFS blocks on those DataNode and keep tracking by using NameNode
* **Secondary** **NameNode** (SNN): is an assistant daemon for monitoring the state of the cluster HDFS. And each cluster had only on SNN.
* **JobTracker**: this daemon work as a connection between the application and Hadoop. Monitors task and assign them node. And also every cluster have one JobTracker
* **TaskTraker**: this daemon does a lot; it responsible: supervision the MapReduce job, manage TaskTrackers, manage task execution on each slave node[[[15]](#endnote-15)].

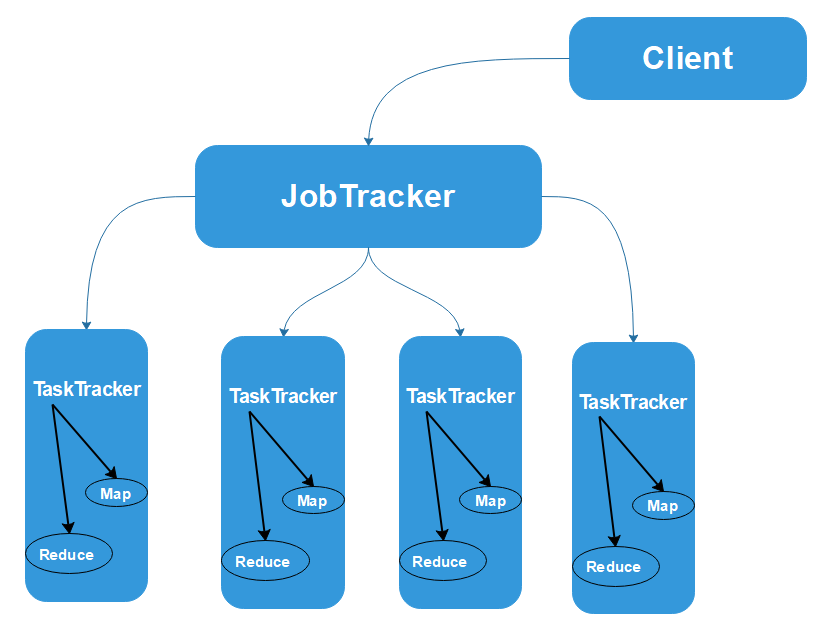


Figure 2.4 Interaction between JobTracker and TaskTracker

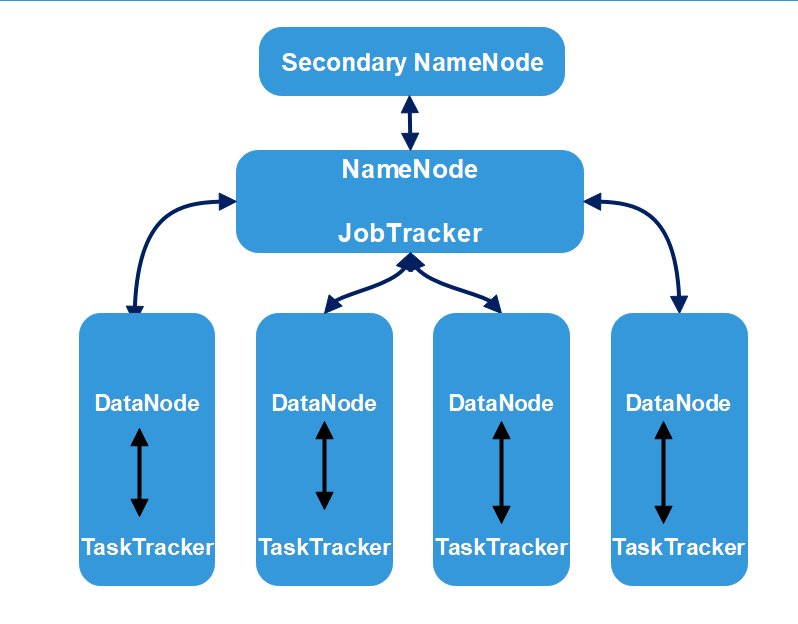


Figure 2.5 The topology of a Hadoop cluster

2.3.2 Hadoop Ecosystem

Hadoop ecosystem is a collection of tools and systems that run with the Hadoop core. It's changed over time and along with distribution. The famous distribution includes Cloudera, Hortonworks, and MapR. Those tools and systems distributed between data ingestion, workflow, message manager, analyzing, and other tasks. Tools and systems like HBase, Spark, Hive, Kafka, Storm, Flume, ZooKeeper, and uncountable other. [[[16]](#endnote-16)]

Hadoop ecosystem aim to increase the accessibility of Hadoop and distributed between high-level languages(ex. Pig, Crunch), SQL-on-Hadoop (ex. Hive, Impala), Predictive analytics (ex. Mahout, R), alternative processing (ex. Storm, Spark), and miscellaneous (ex. Sqoop, Oozie)

**1. Hive**

Apache Hive is a data warehouse framework built on Hadoop. It provides us with the ability to read, write, and manage large datasets. Apache Hive allows you to write SQL queries to retrieve data stored in different databases and file systems. Apache Hive enables data storage tasks such as extract/transform/load (ETL), Hive provides an SQL language, called Hive Query Language (HiveQL or just HQL) for querying data stored in a Hadoop cluster. It also offers direct access to files stored in Apache HDFS or other storage media such as Apache HBase. Apache Hive was initially used by Facebook but has also been certified by other companies. In data warehouse applications it is appropriate to use Hive, where the data that is slowly changing is analyzed, and a rapid response is not required. Because Hive is not a full database, it had its limitations. In Hive, we can not perform a record-level update, insert, or delete. Hive does not provide the fundamental capabilities required for OLTP and online transaction processing. It is closer to being an OLAP tool, online analytical processing. [[[17]](#endnote-17)] [[[18]](#endnote-18)**]**.

The major components of the Apache Hive architecture can be described as follow:

* **Hive metastore** or referred to as HCatalog also, stores metadata such as column names, data types, comments, etc. in table-like structure. Hive provides three main data structures: tables, partitions, and buckets. Hive stores the data in a traditional Relational Database Management System (RDBMS) format. The tables correspond to HDFS directories and may be divided into partitions, where data files can be divided into buckets. Hive's metadata structure is the schema of the Schema-on-Read concept on Hadoop, which implies you are doing not should define the schema in Hive before you store data in HDFS. Applying Hive metadata after storing data brings more flexibility and efficiency to your data work. The recognition of the Hive's metadata makes it the standard way to describe big data and is employed by many tools within the big data ecosystem. The figure 2.6 show the structure of Hive and its relation with Hadoop core.

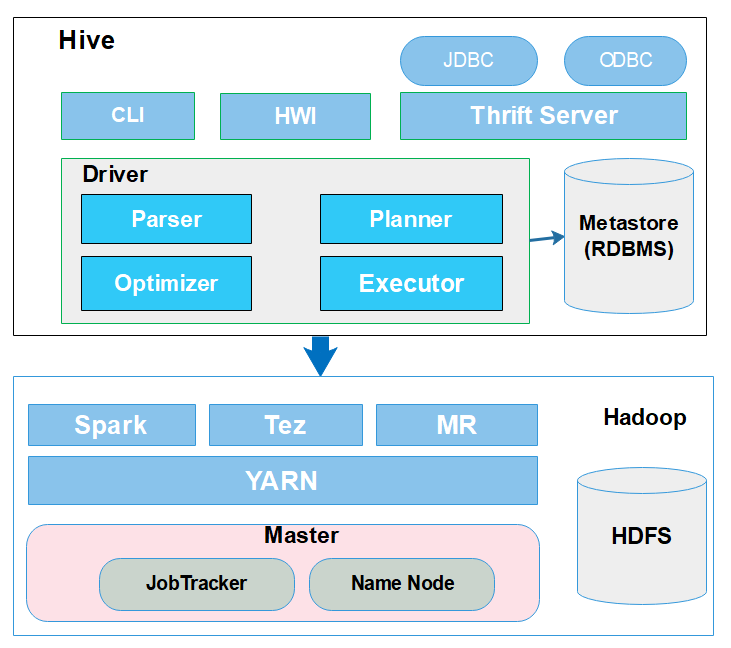


Figure 2.6 the structure of Hive

* **Driver**: in Apache Hive driver acts sort of a controller. Which is to blame for receiving the queries submitted by thrift, Java Database Connectivity(JDBC), Open Database Connectivity(ODBC), command Interface(CLI), Web User Interface(Web UI )which are a part of the Hive client. The driving force stores the mandatory metadata generated during the execution of a HiveQL statement. The driving force interacts with the compiler to urge a thought to perform the execution of a question and collects its related metadata information.
* **Compiler:** it is responsible for compiling the HiveQL query by creating a plan for a query to be executed. The plan shaped by the compiler contains the tasks and steps required to be performed by Map Reduce. The compiler communicates with the Meta store to get the metadata request.
* **Execution Engine** is the component responsible for executing the execution plan after compilation and optimization. It schedules how tasks are to be run by interacting with the Hadoop Job Tracker.
* **Hive UI** enables external users to interact with Hive by running queries to fetch the result, submit instructions, and performing other processes. However, the Hive Web Interface is a Graphical User Interface, which is an alternative to the Hive Command Line Interface (CLI).

1. **Apache NiFi**

For developing the data integration layer, Apache Nifi had been used. Apache NiFi is designed to automate data migrate from one external system to another. In Apache Nifi, "data flow" is used to mean the automated and managed flow of information between systems [[[19]](#endnote-19)].

Apache Nifi is a web-based tool that offers an experience between the design, control, feedback, and monitoring of a data ingestion platform. We can say it is easy to configure and confirms data ingestion, low latency, high throughput, and enables tasks to be dynamically prioritized. Flows can be modified at runtime without any trouble in Apache Nifi. Data provenance service is provided which allows tracking dataflow from the beginning to the end. It facilitates the creation of processors to ingest data from external sources for developers and performs effective testing with this tool. It is also secure as it supports SSL, SSH, HTTPS, encrypted content, and provides authorization for multiple tenants, internal authorization/policy management.

**Flow** **File**: Represents each object moving through the system. File flow is a single piece of data. It is made up of two components: flow file attributes and flows file content. The flow file content is the data, and the flow file attribute provides information about the data.

**Processors**: The Nifi component responsible for listening for incoming data, pull data from external sources, publish data to external sources, and route, transform or extract information from flow files.

**Relationships**: An association between processors. A processor transfers the flow file to one of the relationships after it has finished processing it. The Data Flow manager then connects each of these relationships to other components to specify where the flow file should go next.

**Controller Service:** the controller service automatically starts when NiFi starts up. It also provides information necessary for other components to use. They are being added and configured by the Data Flow Manager.

**Process Group**: are used to group multiple components. It is preferable to use in cases when a data flow becomes too complicated. The figure 2.7 show the process group, when figure 2.8 show the relation between processors.



Figure 2.7 the process group in Apache NiFi

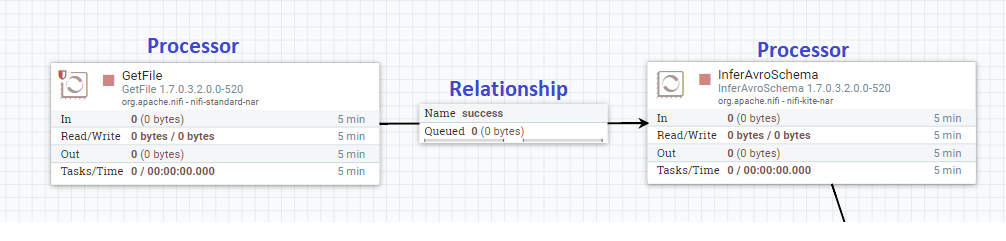


Figure 2.8 processor relationship in Apache NiFi

**Port**: to connect two or more process groups, we need to use ports. Incoming flow files into a process group use input ports, and outgoing flow files from a process group use output ports.

1. **Apache Spark SQL**

Spark SQL is one of the most powerful components of the Apache Spark that has been included since the release of Spark 1.0. It provides support for Java, Scala, Python and the R API. In the figure 2.7, Spark SQL components are shown that support for the application of machine learning, graph, streaming and other applications

Spark SQL is one of the most superior elements of Apache Spark. It has been a part of the core distribution since Spark 1.0 and supports Python, Scala, Java, and R programming APIs. As illustrated in the figure 2.7 Spark SQL components supply the foundation for Spark machine learning applications, streaming applications, graph applications, and many other types of software architectures.

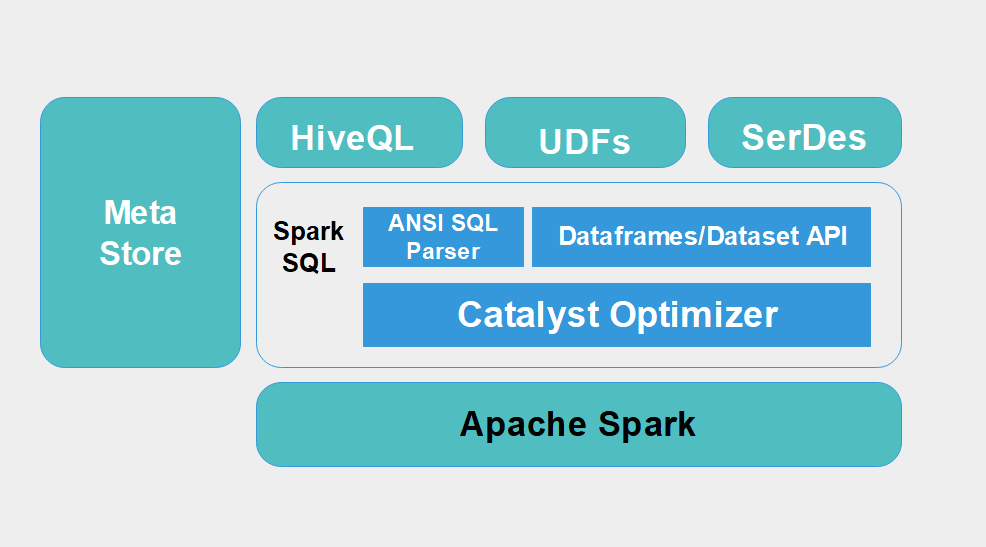


Figure 2.7 Apache Spark and Apache Spark SQL

Apache Spark SQL enables us to query structured data inside Spark programs. It provides a programming abstraction called DataFrame, and it can provide an SQL engine to ensure distributed processing. Apache SparkSQL had the ability to perform in-memory processing, which makes it very suitable to perform analysis. It is faster than HiveQL by 100 times.

Apache Spark SQL is integrated, provides unified access, easily integrates with HIVE, provides standard connectivity, and scalable.[[[20]](#endnote-20)] [[[21]](#endnote-21)**][[[22]](#endnote-22)**]

1. **Apache Zeppelin**

**Apache Zeppelin** is a web-based notebook that enables its users to create interactive data analytics. Apache Zeppelin also allows data ingestion, discovery, visualization, sharing, and collaboration with Apache Spark, SQL, Scala, etc.

With Zeppelin, we can make beautiful data-driven, interactive, and collaborative documents with a rich set of pre-built languages: Python, SparkSQL, Hive, Markdown, Angular, and Shell.

Apache Zeppelin provides a JDBC interpreter, which enables you to create a JDBC connection to any data sources smoothly and perform SQL queries such as Insert, Updates, etc. It has been tested with PostgreSQL, MySQL, Maria DB, Redshift, Apache Hive, Apache Phoenix, Apache Drill, and Apache Tajo. Apache Zeppelin provides a whole range of a group of interpreters for Apache Spark interpreter. It also supports interpreters for Python [[[23]](#endnote-23)]

2.3.3 Cloudera Distributed Hadoop (CDH)

CDH is Cloudera's one hundred percent open source platform distribution, which includes Apache Hadoop and constructed specifically to meet organization demands. CDH provides the entirety you want for company use proper out of the box. By integrating Hadoop with greater than a dozen different imperative open source projects, Cloudera has created a functionally advanced device that helps you perform end-to-end Big Data workflows. Here we can install Apache Hadoop core, Apache Accumulo, Apache HBase, Apache Hive, Apache Flume, HUE, Apache Impala, Apache Kafka, Apache Pig, Apache Sentry, Apache Sqoop, and Apache Spark. in addition, you can install the component of Cloudera Flow Management (CFM)

CDH is the most complete, tested, and famous distribution of Apache Hadoop and associated projects. CDH gives you the core factors of Hadoop – scalable storage and allotted computing – along with a Web-based user interface and necessary agency capabilities. CDH is Apache-licensed open supply and is the only Hadoop solution to offer unified batch processing, interactive SQL and interactive search, and role-based get entry to controls. The figure 2.8 show the typical architecture for Hadoop distribution. [[[24]](#endnote-24)][[[25]](#endnote-25)][[[26]](#endnote-26)]

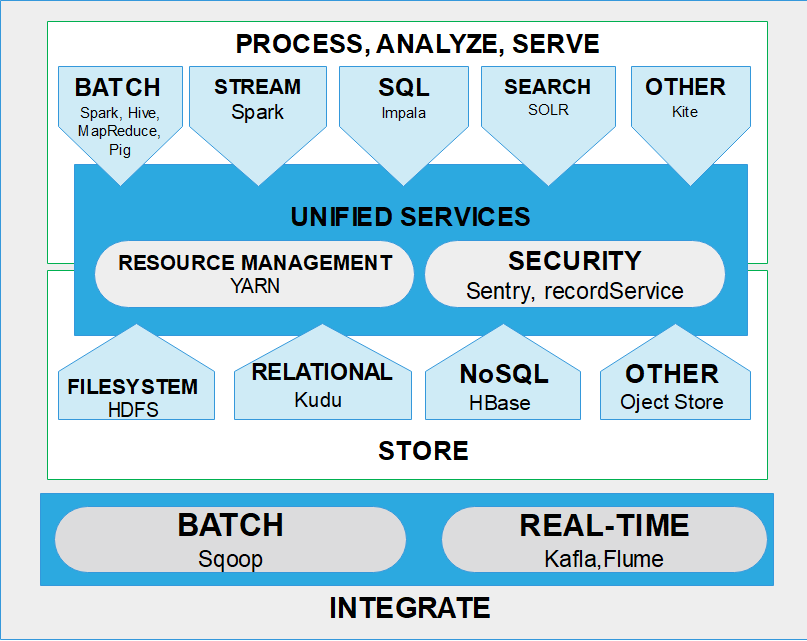


Figure 2.8 Hadoop distribution integrated hierarchy

## 2.4 Hadoop for agriculture

With all those benefits Hadoop provides, it becomes widely used in many sectors and agriculture is not far from that. Agriculture must now feed 8.5 billion people by 2030 with many environmental and ecological challenges. Digital agriculture was introduced to increase yields and reduce losses and wastage. With improved supply chains, it relies on high-quality data to improve decision-making. Providing this data on various aspects of the agricultural process and analyzing it will help support this solution to help provide food.

2.4.1 The big data benefit in Agriculture sector

Big data will provide many benefits for digital farming. Using it, we can benefit from artificial intelligence and machine learning, in addition to using the Internet of Things, which helps to:

* **Make farms smarter** (driverless tractors, drones, smart irrigation, smart greenhouses)
* Interpretation of previous hardware data and building models and prediction
* Using machine vision to diagnose pests and soil defects

Big data ecosystem (Figure 2.9) includes data providers( data brokers and suppliers), big data service providers and big data service customers

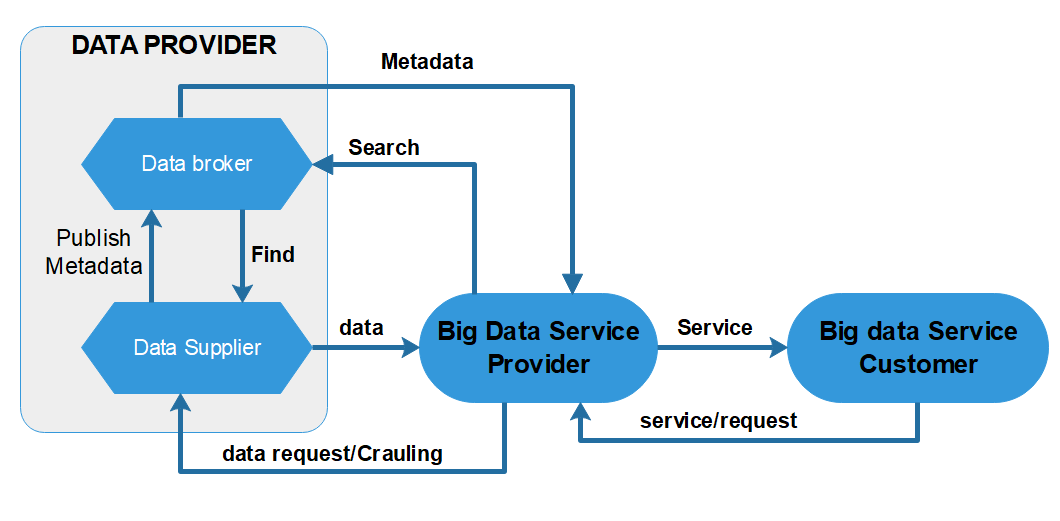


Figure 2.9 Big Data Services Providers, and Customers

In this diagram, the suppliers provide data from various sources to the brokers. Then the brokers work as a connector between suppliers and bigdata service providers. The big data services provider is the big data platform that gives us the power to analyze data besides the infrastructure. The last part is customers or the end-user, which consumes the services from the big data services provider.

2.4.2 Big Data in Agriculture Use Case

1. **OFIS**

Agriculture communities benefit big data in many ways. Like Olam Farmer Information System (OFIS), which use for improving smallholder productivity livelihoods through allowing field staff to survey and record on the spot. That let them get better advice on interventions and compare progress on these as well as identify hotspot for risks, OFIS also gives insights to tackle issues ranging from poor yields to climate change.

OFIS start in 2014 at Cote d'Ivoire to understand the farm landscape of Olam's cocoa suppliers. But since then has been rolled out to eight product categories: coffee, cashew, cotton, and rice, in 27 countries.

It collects data about farm size, location, age of tree stock, economic, social, and health infrastructure, and eco-support system using a handheld device.

1. **mSTAR**

data scientists are piloting new methods to estimate yields, growing areas, and even poverty, and research institutions are demonstrating what can be gained by analyzing large amounts of data on crop management, yield, soils, and weather conditions in local farming systems.

mSTAR project supported the Digital Development for Feed the Future  
(D2FTF) a team within USAID's Center for Digital Development, USAID/Cambodia, and USAID/  
Nepal over the course of 14 months to identify ways to improve the structure, storage, and  
governance of FTF data and facilitate analysis across the portfolios[[[27]](#endnote-27)].

1. **CGIAR**

Their goal, according to them, Using big data approaches to solve agricultural development problems faster, better, and at a greater scale than before.

The CGIAR platform for the Big Data is implemented on farms with the support of the Development Fund with a bilateral financing agreement. The aim of the platform is to take advantage of the data capabilities available to accelerate and enhance agricultural research worldwide. The platform duration is five years 2017 – 2021 and it provides open data and organizes it. What is required from the partners is to provide innovative ideas and verify the power of analyzes provided by big data.[[[28]](#endnote-28)]

There are many success cases in developing countries, for example in Kenya, which is a distinct center of information technology innovations for farmers, where they can access the extension records provided by iCow, which change the concepts and methods of agricultural extension. Also in Mexico, MASAGRO, which the Mexican government operates with CIMMYT, has developed an application to provide satellite images to help farmers calculate the amount of nitrogen their farms need. [[[29]](#endnote-29)]

2.4.3 What is Big data can offer for agriculture:

**Business solutions**

The big data in agriculture thus collected from such interactions regarding crops, weather, terrain, geographic conditions, water, and more are stored and processed. This leads to the analytics part of our solution. By processing this data, the application will be able to assist:

* Agriculture companies regarding the prospective success of products in different markets
* Farmers about the success of various crops, the predictive impact of natural conditions, etc.
* Consultants will learn more about the most affected geographic area where farmers could be assisted to deliver higher productivity.

**Technology** **solutions**

Such applications developed with Hadoop and any front-end technology using deep learning/machine learning allow the platform to provide:

* Predictive analytics: to predict the success of product or crop or predict the ill effect of a natural event on crops
* Historical data analysis: process vast volumes of historical data regarding crops, geography, etc
* Real-time analytics: to provide farmers with real-time assistance by analyzing information provided by them in real-time

**Business benefit**

The benefit expected from such application as follows

* Increase agricultural productivity: through agriculture data
* The great success of fertilizing products across a variety of geographic conditions for agriculture companies

# 3 Data Recognition

## The current system in the Planning and Agricultural Economics

The Department of Agricultural Planning and Economics consists of five administration. In each administration, a partial system is used to deal with data, as most of it handles data in Excel format or separated values. Some also use regular text files. In addition to the presence of a database using MySQL, but this distribution leads to an impact on the efficiency of analyzes in management as a whole.

3.1.1 The Text files

The department receives an amount of information from several ministries and other departments in the form of text files, whether in Excel or CSV format. The department is located in the city of Khartoum, and it receives these data from departments and ministries located in the same city, with branches in different cities of Sudan. These ministries receive their information from their branches on weekly and monthly grounds while it provides the Department of Agricultural Planning and Economy on a monthly bases.

These data are collected monthly and prepared to generate reports for the administration. The agricultural statistics administration of the department supervises this. We discussed the data sources in the first chapter in our discussion of background according to the figure 3.1

Figure 3.1 diagram showing PAE reporting flow

3.1.2 The MySQL Database

The Department also has a database that uses MySQL, which collects this information from several sources in cooperation with the relevant departments. These data are then provided to the Agricultural Statistics administration and other offices that need reports. The task of managing and feeding this database depends on the Information Technology Administration of the Department.

Figure 3.2 Reporting flow for MySQL database

3.1.3 The weather Data

The administration also receives weather data from the national meteorology Authority. Here we preferred to receive data from an international weather website using the API for the accuracy and speed of the data it provides if compared to the infrastructure owned by the Sudan Meteorological Authority.

## Data structure

* + 1. CSV data structure

We chose four of the primary crops that can be considered as cash crops for Sudan, namely: wheat, sorghum, millet, and cotton because Wheat, Millet, and Sorghum are introduced as a fundamental component of food in Sudan, mainly wheat in the north and cities, while millet is more concentrated in the west, and it and sorghum share consumption in all Sudan to varying degrees. All values ​​are for the years from 1960 to 2019. Some properties may differ depending on the type of crop. The basic features of these crops are as follows:

* + - 1. **Harvested Area**: it is Indicates for the area from which the crop is collected. It is the harvested area only, as the uncultivated or un harvested area is excluded, and also the unharvested area without cause or due to damage are exclude. [[[30]](#endnote-30)]
      2. **Beginning Stocks**: Existing supplies of a crops commodity that consist of remaining stock carried over from the previous year's production. [[[31]](#endnote-31)]  
         the ending stocks carried into the new marketing year from the previous year
      3. **Domestic Consumption:** the quantity that is harvested or imported and used in the country[[[32]](#endnote-32)]. So it contents of all possible uses of the commodity: food, feed, seed, waste, and industrial processing
      4. **Ending Stocks**: The remainder of current crop production carried over into the next crop year. The unused commodity was remaining at the end of the marketing year for use in the next year.
      5. **Exports**: The quantity transferred from the crop to another country or region for trade purposes
      6. **Imports**: The quantity brought from another country for use within Sudan
      7. **Production**: It is the actual harvested output of the crop, with the exception of the harvest and threshold losses, and the portion that is not harvested for any reason. Therefore, production includes both the quantity sold in the market and consumed by the producer. If the production data refer to two calendar years and it is not possible to divide the production between them, then we attribute it to the year in which the largest amount of production occurred. We use the metric ton to calculate the output (MT).
      8. **Total** **Supply**: mainly it calculated by the following formula  
         Total Supply = beginning stocks + domestic production + imports
      9. **Yield**: The data represents the harvested output per unit. This value is usually calculated by dividing the production by the harvested area. Data are recorded in a hectare (100 kilograms per hectare) (KG/HA).

For each of the above properties, we will study the following features

* **Market** **Year**: It is the year in which the production process took place
* **Quantity**: it varies according to the property. If we were studying, for example, the export, then it would be the exported quantity
* **Unit** **of** **Measure**: The unit of measurement varies according to the crop and the property; for example, with cotton will be Bales, which is equal 480 lb. With the harvested area, it will be one thousand hectares, which equals 10 square kilometers, and it is shown in the table 3.1

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | equivalent in the metric |  | Measure unit | property |
| Cotton | cereals | Cotton | Cereals |  |
| 0.01 km2 | 0.01 km2 | HA | HA | Area Harvested |
| 217.724 KG | 1000 KG | 480 lb. Bales | MT | Beginning Stocks |
| 217.724 KG | 1000 KG | 480 lb. Bales | MT | Domestic Consumption |
| 217.724 KG | 1000 KG | 480 lb. Bales | MT | Ending Stocks |
| 217.724 KG | 1000 KG | 480 lb. Bales | MT | Exports |
| 217.724 KG | 1000 KG | 480 lb. Bales | MT | Imports |
| 217.724 KG | 1000 KG | 480 lb. Bales | MT | Production |
| 217.724 KG | 1000 KG | 480 lb. Bales | MT | Total Supply |
| KG/KM2 | KG/KM2 | KG/HA | MT/HA | Yield |

Table 3.1 The measures unit for Cereals and cotton and equivalent metric

* **Growth** **Rate**: Growth rates refer to the percentage change of production within a specific time period. And we calculated it by the formula



Where:

PR = Percent Rate  
VPresent = Present or Future Value  
VPast = Past or Present Value

* + 1. The relational database data structure

The domestic production database is a small database designed using MySQL to collect some essential data. According to the Administration of Agricultural Statistics, they are working on a more comprehensive database, initiated in cooperation with the Information Technology Administration to implement. It is well known that the MySQL database is a widely used database worldwide and supported by almost all hosting providers, was open source, and free before they became Oracle property.

In this research, the following tables will be used from the database that we display the ERD for it in figure 3.3

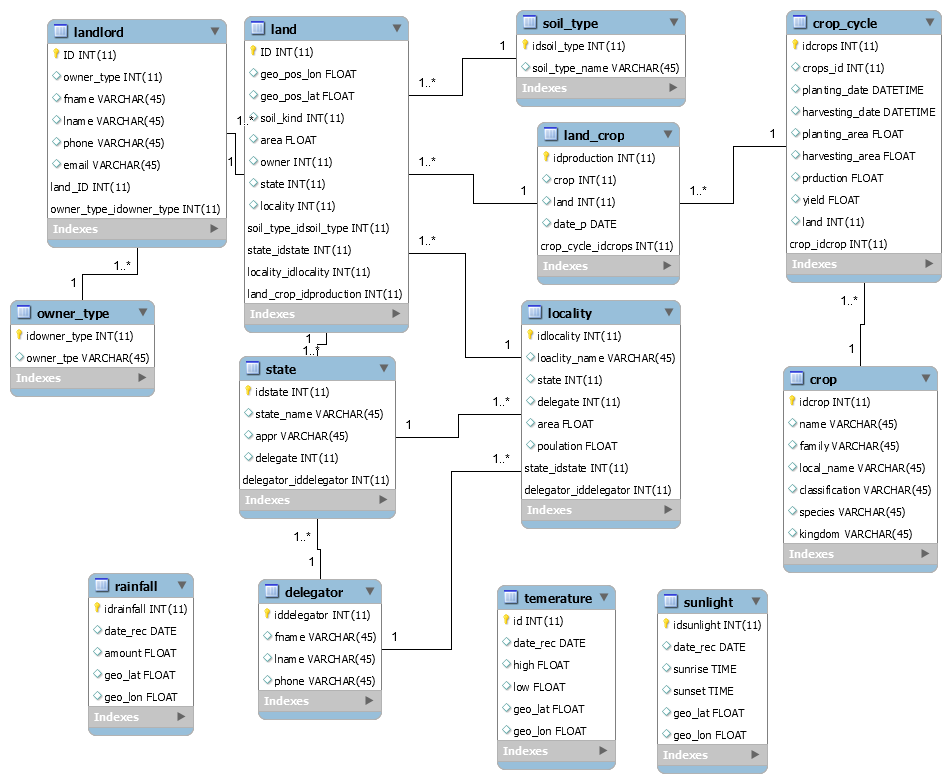


Figure 3.3 Entity relation diagram for PAE\_DB database

The tables for this database are:

1. **landlord** table: In this table, information is provided about landlords and their contact information.
2. **owner\_type** table: This table stores information about the type of landowner who is divided into individuals, companies, and projects owned by the state. In order to help explain the contribution of each sector to domestic production.
3. **Land** table: Contains important information related to the cultivated land, such as its geographical location, area, soil type, landlord, and administrative belonging.
4. Table of **soil\_type**: data stored on this table is the type of soils, according to relation to the diversity of soil types in Sudan, which helps in providing advice to owners on the most suitable crop type for cultivation. It also helps in studying the productivity of each type of land for the specific crop.
5. **crop\_cycle** table: This table stores information related to the agricultural cycle, such as the dates of planting and harvesting, and the cultivated and harvested areas for the farming season. You will use here the same measures mentioned in the table 3.1 which is a hectare and metric ton
6. **crop** table: Here information is stored about the agricultural crop, such as local and scientific name, classification, family, and kingdom. This helps in subsequent scientific research on a specific family or specific crop.
7. **State** table: Important data for Sudan's 17 states: two states in the north, three in Kordofan, five in Darfur, three in the east, and five in the center, including Khartoum state, which contains the country's capital.
8. **Locality** table: Sudan consists of about 189 localities that differ in their areas and their number in each state, as it reaches 21 in some states, while they are only 3 in other states. They also differ in the prevailing climate, from desert to subtropical. This table stores information about the state, its area, and its population, in addition to the delegator.
9. The **Delegator** table: The person responsible for communicating agricultural statistics, in the specific locality or state, to the Department of Planning and Agricultural Economics.
10. **Land\_crop** table: This table is used as a link table to provide us with a many-to-many relationship between the lands and crops tables because it is known that every piece of land owned by an individual or group may produce more than one crop and that a single crop can be grown on more than one land. We also define the irrigation method used, which is divided into two sectors: rain and irrigated.

The following three tables are for recording weather conditions, temperature, amount of rain and sunlight, and the geographical location is also recorded, even if it lacks wind intensity and humidity. It is used to set forecasts for the amount of rain and also helps to predict productivity in the event of favorable weather.

1. **Temperature** table: In this table, the information related to the temperature is stored as the lowest and highest degree with the determination of the location
2. **Rainfall** table: We record data related to the amount of rain, date, and location
3. **Sunlight** Table: It stores sunlight data such as sunrise and sunset.
   * 1. Data structure of Weather API

To get weather data, we use the API method to get our data from OpenWeatherMap. [OpenWeatherMap](http://www.openweathermap.com/) gives current weather information and forecasts. It is very simple to use. The [OpenWeatherMap weather API](http://www.openweathermap.com/api)uses JSON to exchange data.

The weather API returns current weather, forecast, historical weather information, and weather station data. Moreover, it also has UV index data.

OpenWeatherMap provides different two modes to look for a city. One uses name pattern and another using geo-coordinates

All these APIs are free to use, but we have to create a dev key so that you can make requests.

when we call using the following API call

<https://samples.openweathermap.org/data/2.5/weather?id=379252&appid=b6907d289e10d714a6e88b30761fae22>

we get the following result:

{

"coord":

{

"lon":145.77,

"lat":-16.92

},

"weather":

[{

"id":802,

"main":"Clouds",

"description":"scattered clouds",

"icon":"03n"

}],

"base":"stations",

"main":

{

"temp":300.15,

"pressure":1007,

"humidity":74,

"temp\_min":300.15,

"temp\_max":300.15

},

"visibility":10000,

"wind":

{

"speed":3.6,

"deg":160

},

"clouds":

{

"all":40

},

"dt":1485790200,

"sys":

{

"type":1,

"id":8166,

"message":0.2064,

"country":"AU",

"sunrise":1485720272,

"sunset":1485766550

},

"id":2172797,

"name":"Cairns",

"cod":200

}

the responds in JSON file providing the following attributes:

* **coord**
  + **coord**.**lon** City geolocation, longitude
  + **coord**.**lat** City geolocation, latitude
* **weather** (more info Weather condition codes)
  + **weather**.**id** Weather condition id. It divided into seven groups
  + **weather**.**main** it is the group that content weather parameters (for example Rain, Snow, Extreme, etc.)
  + **weather**.**description** Weather conditions. You may prefer to get the description in your language.
  + **weather**.**icon** the id for weather icon
* **base** base present the Internal parameter
* **main**
  + **main**.**temp** the temperature. We can use any unit, but the default one is Kelvin, Metric or Celsius, Imperial, and Fahrenheit.
  + **main**.**feels\_like** for temperature also. It represents the human perception about the weather.
  + **main**.**pressure** this parameter gives us the atmospheric pressure (measure on the sea level), it uses hPa
  + **main**.**humidity**: Humidity, %
  + **main.temp\_min** the amount of minimum temperature.
  + **main.temp\_max** the amount of maximum temperature.
  + **main.sea\_level** it measures the atmospheric pressure that toke on the sea level, it measures using hPa
  + **main.grnd\_level**: it measures the atmospheric pressure that toke on the ground level; it measures using hPa.
* **wind**
  + **wind.speed** parameter for wind speed. The default unit for the metric is meter/sec, and for imperial are miles/hour.
  + **wind.deg** it gives the wind direction.
* **clouds**
  + **clouds.all** Cloudiness, %
* **rain**
  + **rain.1h** it retrieves the rain amount for the last 1 hour, mm
  + **rain.3h** it retrieves the rain amount for the last 3 hour, with millimeter mm
* **snow**
  + **snow.1h** it retrieves the snow amount for the last 1 hour, mm
  + **snow.3h** it retrieves the snow amount for the last 3 hour, with millimeter mm
* **dt** Time of data calculation, UNIX, UTC
* **sys**
  + **sys.type** Internal parameter
  + **sys.id** Internal parameter
  + **sys.message** Internal parameter
  + **sys.country** Country code (GB, JP, etc.)
  + **sys.sunrise** Sunrise time, Unix, UTC
  + **sys.sunset**: Sunset time, Unix, UTC
* **timezone** Shift in seconds from UTC
* **id** City ID
* **name** City name

**cod** Internal parameter

## Data Sources

The primary data source in this research is the data we obtain from the Department of Planning and Agricultural Economics. It was very difficult to obtain for a variety of reasons, most notably the bureaucratic complications and the lack of good infrastructure to store and retrieve data. Where we initially made a questionnaire distributed to the various administration in the department to find out the type of data available in each administration it appears in figure 3.4

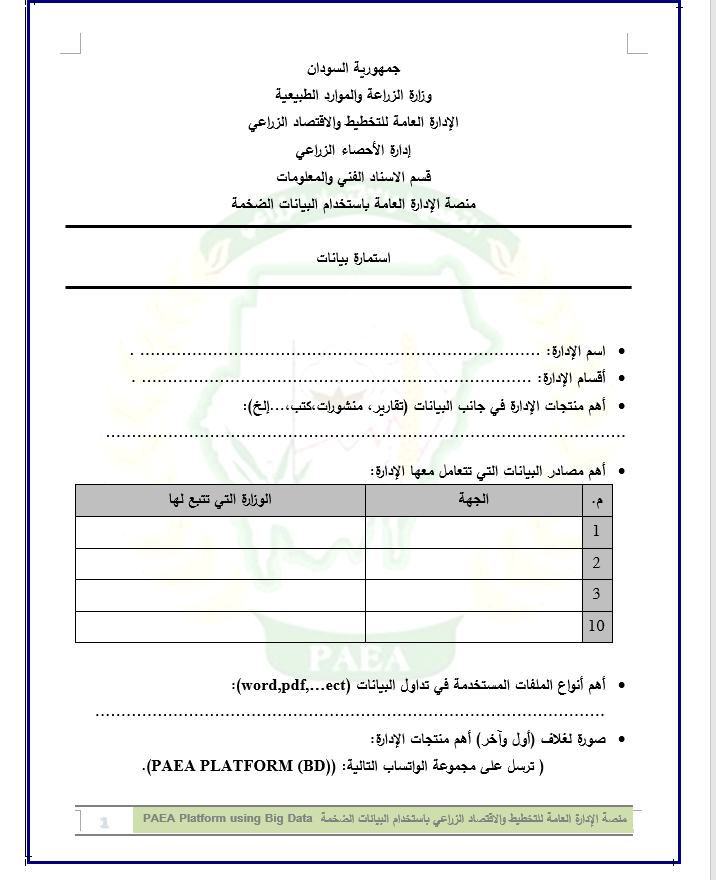


Figure 3.4 Form to determine the data exchange formats

This helps us to know about the type of data available in each administration; most of them were in doc format in addition to Xls with some data in SPSS format. We benefited from it in the first type of data after cleaning the data and converting it to CSV format.

Also, available data in the database, although incomplete in the absence of data for some years.

We had to resort to global databases, so we got our data from the FAO, which offers the FAOSTAT database, which gives you free access to food and agricultural data from 1961. And the good thing about the FAO site is that they also provide the data in CSV format.

As for the weather data, we tried to communicate with the Sudanese Meteorological Authority, but we did not obtain data from them, so we turned to the Internet and obtained access to the openweathermap.com website, as mentioned above. It provides free access to the current weather forecast with some restrictions. With the provision of access to historical data if you are a subscriber.

# System Design and Data Structure.

## 3.1 Data Structure Design

In this study, we want to develop a platform that collects data from multiple sources in the Department of Planning and Agricultural Economics, where it collects the data available in CSV format, MySQL database, and the current weather data from API Stream. The benefit is providing the data necessary to carry out an effective analysis process that helps make wise and more scientific decisions quickly and easily. The platform consists of:

### 4.1.1 Data ingestion

In this layer, we make sure to obtain data from the three different sources to be sent to the storage layer to ensure that it is stored in one warehouse. This process will be done automatically at specified intervals. It performs four functions:

**Data Connector:** This function enables the data integration subsystem to connect to the data sources of the external systems.

**Data Selector**: This function shall be responsible for selecting source data for data integration from different data sources.

**Data Aggregation**: This function shall be responsible for aggregating data selected by the Data Selection function.

**Data Uploader:** Shall be responsible for uploading aggregated data to the main repository.

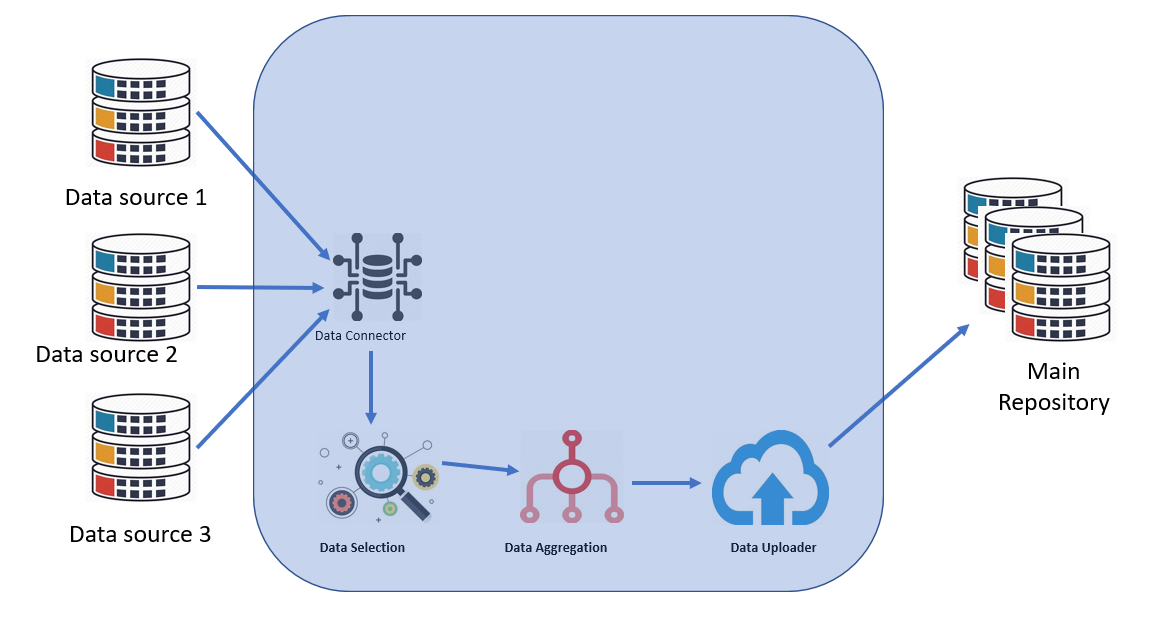


Figure 4.1 data ingestion layer

### 4.1.2. Data storage

The main responsibility for data storage layer is storing data and converting them into formats that are easy for analysis tools to handle. The data will be placed in HDFS with the use of hive tables, which we discussed in Chapter Two. It provides us with data access when needed while ensuring that it is stored. It also converts data into formats understood by analysis tools. So it offers two functions:

**Data conversion function:** This function converts the data in a format that analytical tools can understand.

**Data Storage Function:** Ensure that the data is stored in the main repository.

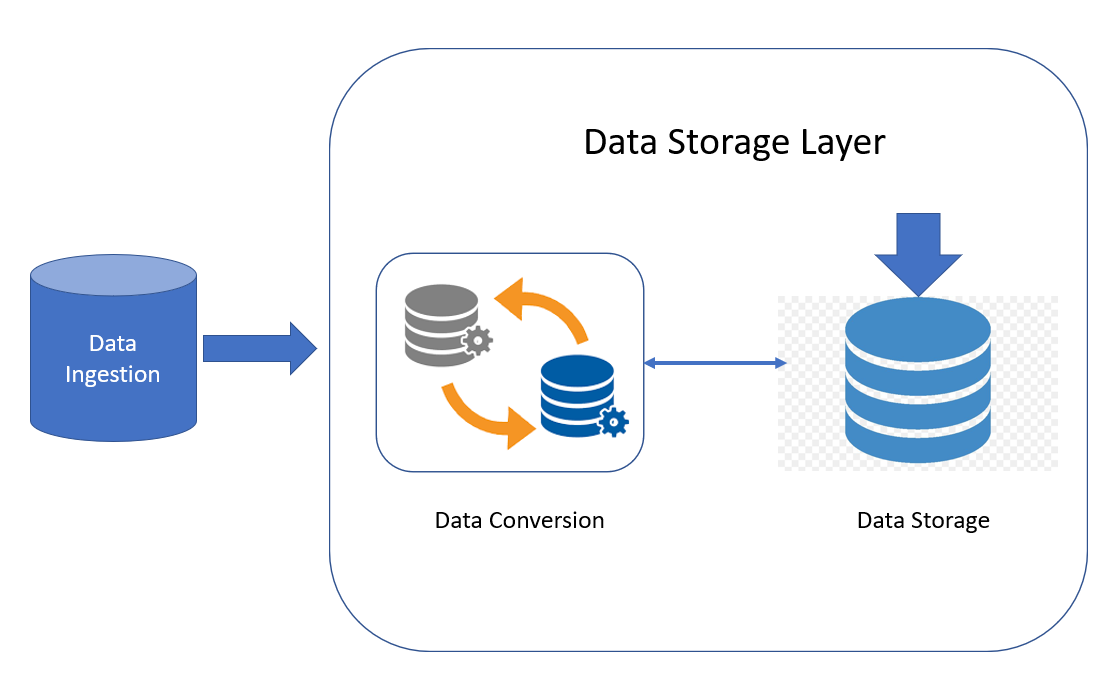


Figure 4.2 Data Storage Layer function

### 4.1.3 Analysis and visualization

We will use in this layer the analysis tools that provided by Hadoop, to analyze the data that we previously stored, and present it in a form, that is easy to understand for the end-user as it is presented in the form of chart and diagrams and this shows the benefit that this platform can contribute. This layer consists of three components

Data Selection Planner:Witch performs three functions.

**Data Connector:** Connects to the data storage layer.

**Data selector:** this function enables one to select data stored in the main repository.

**Data Executor:** This function shall execute the selection plan.

**Data Generator:** This function shall generate a new dataset from the selected data.

#### Data Analysis Component (DAC): This component shall be responsible for performing data analysis on the data ingested to the central repository.

Data Visualization Component (DVC)**:** The data visualization componentshall be responsible for presenting the analyzed data in a form that would be easy to understand by end-users. It has only one function that enables the data to be presented in the form of bar charts, pie charts, histograms, line graphs, etc.

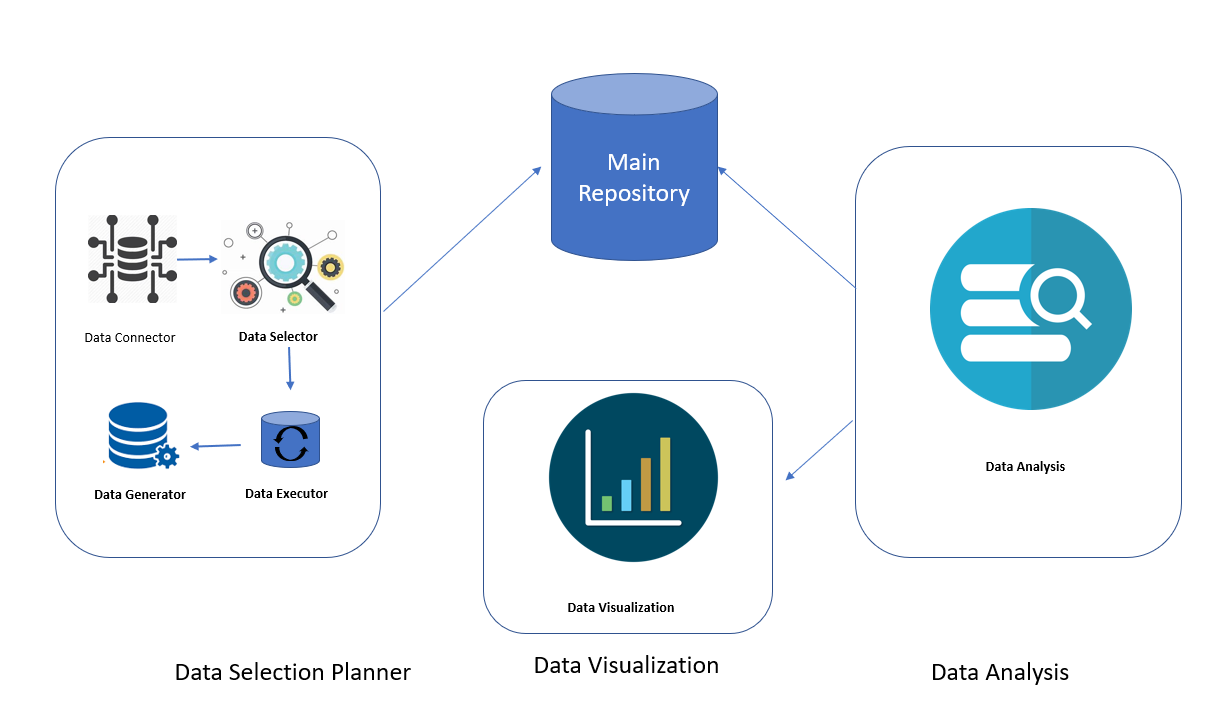


Figure 4.3 Data Analysis and visualization layer

## 4.2 Development Tools

We used different tools from the Hadoop core and Hadoop echo system to build this platform. There are divided on platform layer as follow:

### 4.2.1 Development tools for Data Ingestion Layer

The researcher used Apache Nifi to build the data ingestion layer. Apache NiFi is a tool to migrate data automatically from one external system to another. In Apache Nifi, “data flow” is used to mean the automated and managed flow of information between systems [[[33]](#endnote-33)].

Apache Nifi is a web-based tool that provides an experience between the design, control, feedback, and monitoring of a data ingestion platform. It is easy to configure and ensures guaranteed data ingestion, low latency, high throughput, and enables tasks to be dynamically prioritized. With Apache Nifi, flows can be modified at runtime without any disruption. It provides data provenance service, which allows dataflow to be tracked from the beginning to the end. Developers can rapidly build processors to ingest data from external sources and performs effective testing with this tool. It has very little security vulnerabilities as it supports SSL, SSH, HTTPS, encrypted content, and provides authorization for multiple tenants, internal authorization/policy management.

We discuss Apache NiFi in chapter two. But we mentioned it here as a tools for ingestion data.

### 4.2.2 Development tools for the Data Storage layer

We use the following tools are used to build the Data Storage layer

#### Apache Hadoop

Apache Hadoop is a software framework used for distributed storage and processing of enormous data sets on different clusters of computers built from hardware that are affordable and easy to obtain. It is, however, open source. Hadoop is built with the notion that hardware failures are common things that occur, and the framework should have the capability to handle such. Apache Hadoop enables data storage, data processing, data access, data governance, security, and operations.[[[34]](#endnote-34)]

We speak about apache Hadoop in detail in chapter 2. So here we just want to mention which part of apache we will use in this layer.

#### HDFS

is used to store the central part of data to ensure that our data are stored in distributed manner.

MapReduce: makes it easy to write applications that are capable of processing a massive amount of data in parallel and on large clusters of affordable hardware in a reliable, fault-tolerant manner.

#### Apache Hive

as we said in chapter 2, it is a data warehouse software framework built on top of Hadoop

### 4.2.3 Development tool for the Data Analysis and Visualization layer

#### Apache Spark SQL

Apache Spark SQL and Apache Hive are used to perform data analysis on the stored data in the data storage layer. Apache Spark SQL is integrated, provides unified access, easily integrates with HIVE, provides standard connectivity and scalable

#### Apache Zeppelin

With Apache Zeppelin, Apache SparkSQL is configured to connect to the Hive warehouse, execute queries, and visualize the result of the analysis. HiveQL is also set on this editor to connect to the HIVE warehouse and execute HiveQL queries.

## 4.2 System implementation

To implement the platform according to the requirement. We use the following tools, as we mention in the data structures design part.

Apache Nifi is used in the development of the Data Integration Subsystem. Apache Hadoop and Apache Hive are used for the Data Storage Subsystem. Finally, Apache Hive and Apache Spark SQL are used for the Data Analysis and Visualization Subsystem. Apache Hadoop HDFS, Apache Hive, Apache Spark services are installed on the centos Operating system cluster of 5 nodes using Clodera Manager (CDH). We use a virtual machine to build a cluster consists of five machines: one master node and four slaves node.

### 4.2.1 Implementing the base machine

To download, install, and configure Centos Operating System, We download centos DVD iso 7.2 build no 1511. Install it on the virtual machine with the following parameter

Two GB of RAM

80gb HDD

Two processor

**Building the cluster**

Implementing the platform on a cluster of five node and we use the following configuration as it shown in table 4.1

|  |  |  |
| --- | --- | --- |
| Machine name | Machine IP | Type |
| Mgt01 | 10.0.2.100 | Master Node |
| Dn01 | 10.0.2.101 | Slave Node |
| Dn02 | 10.0.2.102 | Slave Node |
| Dn03 | 10.0.2.103 | Slave Node |
| Dn04 | 10.0.2.104 | Slave Node |

Table 4.1 hostname and IP address for cluster node

So we make a full clone for the base machine. And to change network configuration we modify /etc/sysconfig/network-scripts/ifcfg-enp0s3 as following:

TYPE=Ethernet

BOOTPROTO=static

DEFROUTE=yes

PEERDNS=yes

PEERRPUTES=yes

IPV4\_FAILOURE\_FATAL=no

NAME=enp0s3

DEVICE=enp0s3

ONBOOT=yes

IPADDR=10.0.2.100

DNS1=8.8.8.8

DNS2=8.8.4.4

GATEWAY=10.0.2.1

NETMASK=255.255.255.0

**Disable and stop firewall**

systemctl disable firewalld

systemctl stop firewalld

**Disable selinux**

vi /etc/selinux/config

change selinux to selinux=disabled

**Set hostname**

We use nmtui command to set host name

**Configure NAT Network**

We make configuration on /etc/host, this will make the machine see each other

#vi /etc/host

10.0.2.100 mgt01.cloudera mgt01

10.0.2.101 dnt01.cloudera dn01

10.0.2.101 dnt02.cloudera dn02

10.0.2.101 dnt03.cloudera dn03

10.0.2.101 dnt04.cloudera dn04

#service network restart

Then we clone mgt01 with reinitialize mac address and make the necessary change for hostname and IP address

**Installing Cloudera Manager**

First, we have to change the RAM capacity for mgt01 from 2 GB to 8gb with two core. Then we obtained the Cloudera Manager from the following URL:

# wget https://archive.cloudera.com/cm5/installer/latest/cloudera-manager-installer.bin

# chmod u+x cloudera-manager-installer.bin

# ./cloudera-manager-installer.bin

**Connecting the cluster**

To connect the cluster, we need to enable port forwarding from NAT network.

**Deploying the cluster**

To reach the master node in the purpose of deploying Hadoop to all nodes we use our host web browser to connect using the following URL:

localhost: port

Will start a wizard with the following screen

1. End-User license
2. Cloudera edition. we will select express addition
3. Specify the host: to spacifing the node that we want to involving in this deploying, we will use the pattern :
   1. dn01
   2. dn02.cloudera
   3. dn[03-04]
   4. 10.0.2.100
4. Select the repository
   1. Use Parcels as a method
   2. 5.9.0 for CDH
   3. Match release for this Cloudera manager server
5. Select JDK Installation Options
6. We won’t Enable single-user mode for security reasons. it makes one user control all services
7. Provide SSH login credentials; there are tow way eather providing ssh key or by using user name and password and that what we will going to chose.
8. Cluster installation starts. It consists of many processes in order to deploy it on all five node.

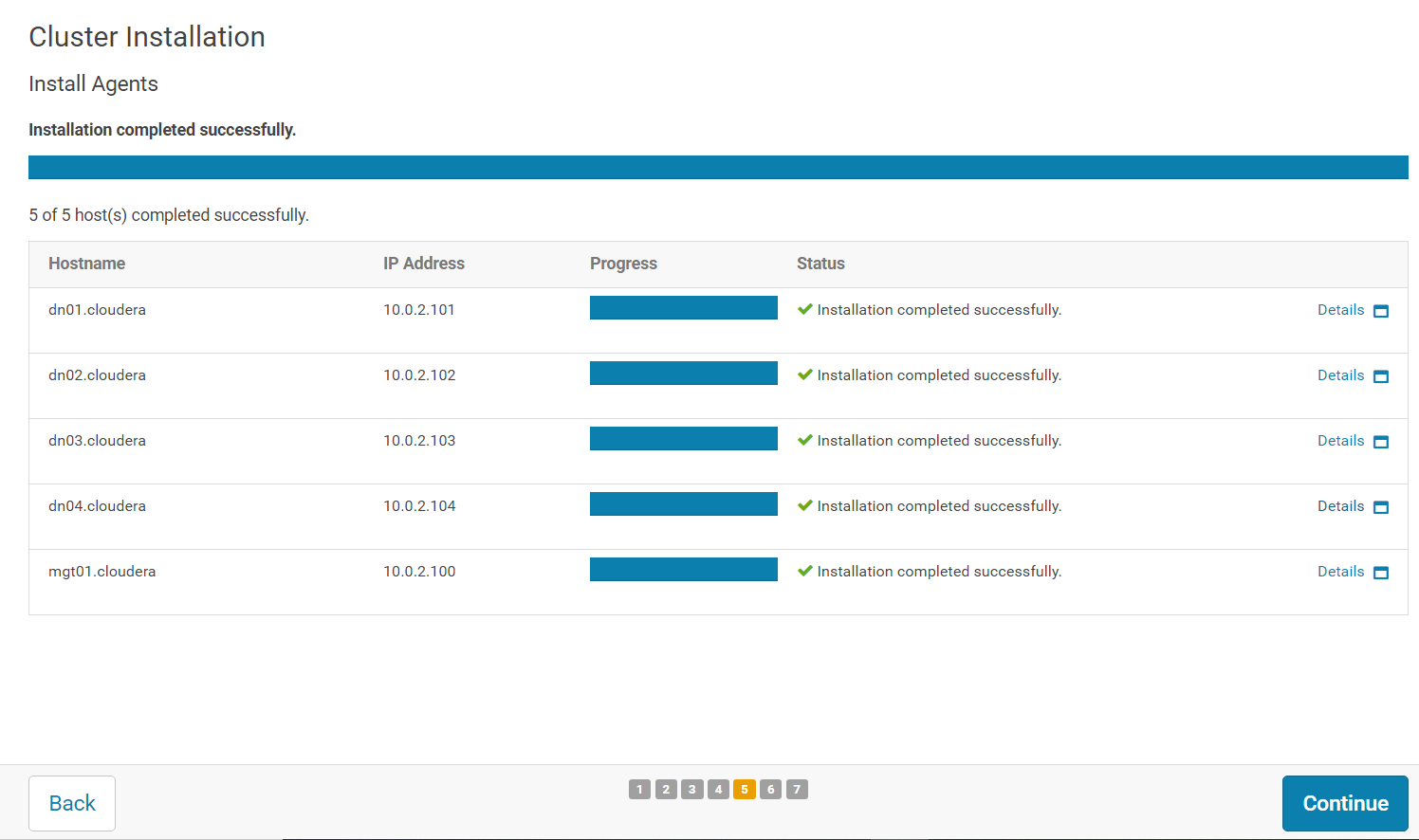


Figure 4.4 Installing Agent in Cluster

1. In the last step, we only deploy the agent, and now we need to deploy the services. So Cloudera manager in this step will start downloading, distributing, unpackaging and activating the parcels we need for our cluster.



Figure 4.5 Installing Parcel step (Downloading, Distributing, Unpacking, and Activating)

1. CM starts to inspect hosts for the correctness. it Will show us a summary with any suggestion needed

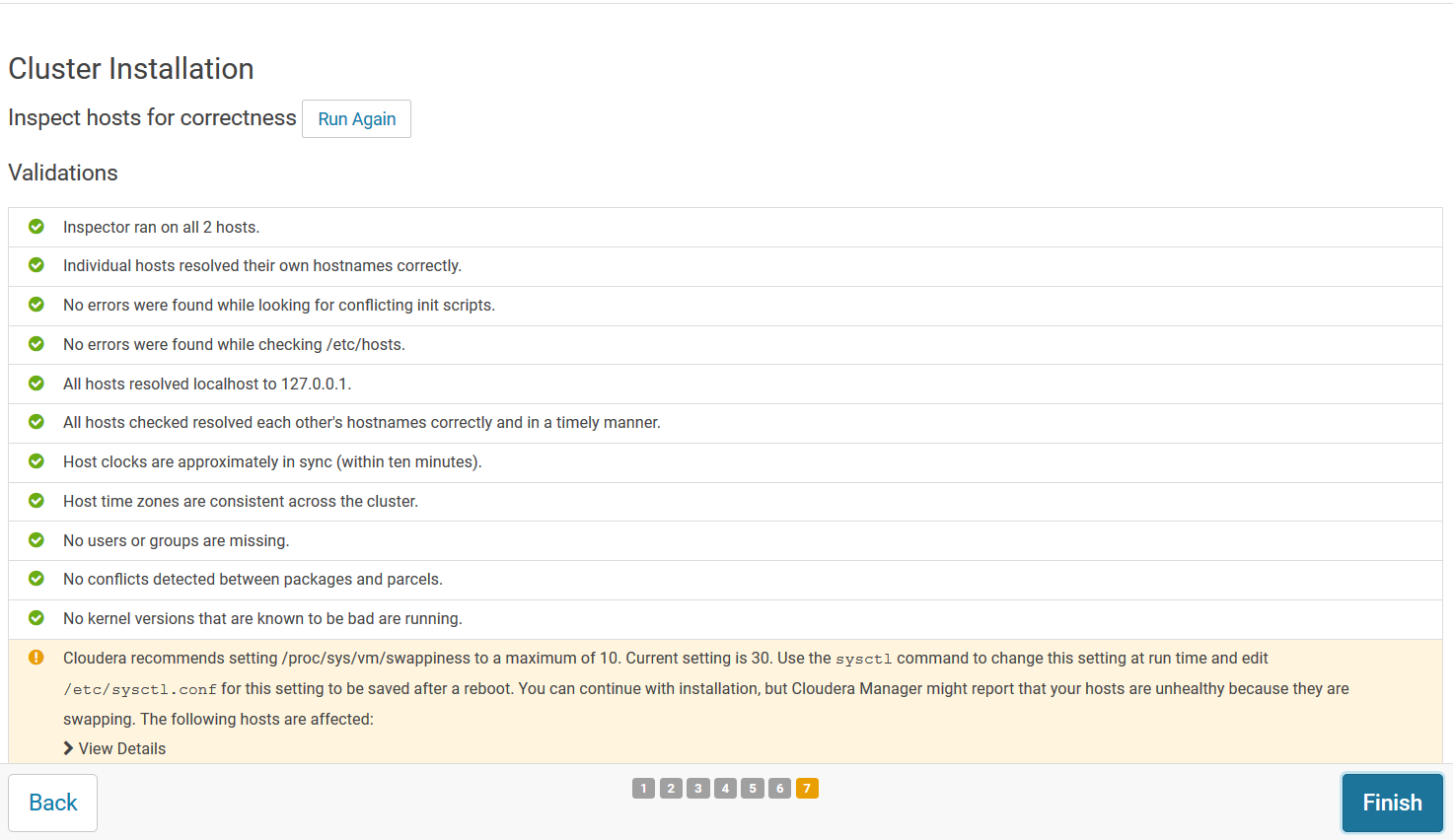


Figure 4.6 Finishing installation of the cluster and inspect host for the correctness

1. After finished the all previous step then we selected CDH services that we need, we started with just Core Hadoop which consist of HDFS, YARN included MapReduce 2, Zookeeper, Oozie, Hive, and Hue.

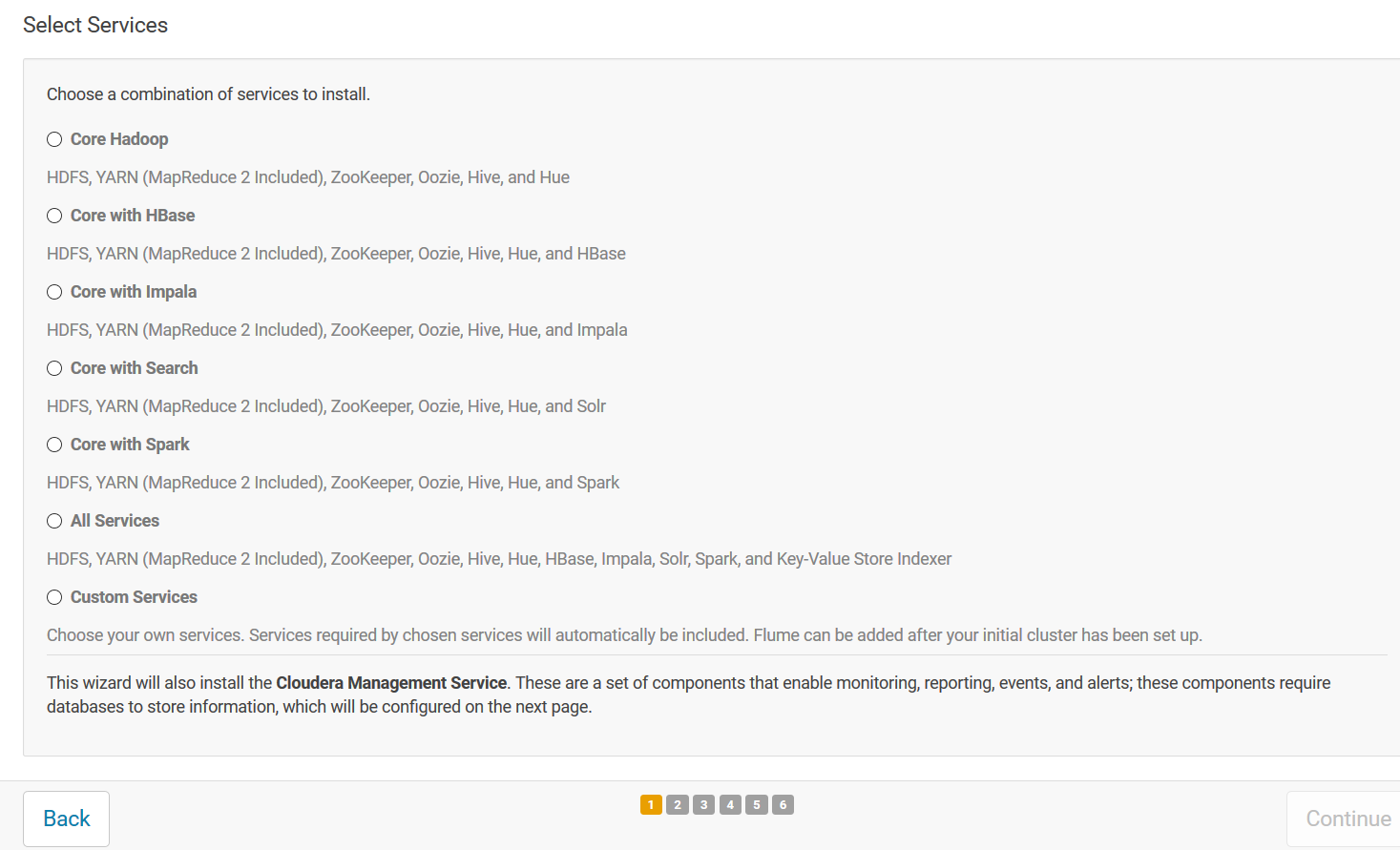


Figure 4.7 Select the type of services to install

1. Customize Role Assignments: to specify the role that we want to assige for the new clusters. We configure it by the following settings as shown in table 4.2

|  |  |  |
| --- | --- | --- |
| Services | Node |  |
| HDFS | mgt01 |  |
| SecondaryNameNode | mgt01 |  |
| Balancer | mgt01 |  |
| HttpFS | mgt01 |  |
| DataNode | mgt01 |  |
| NFS Gateway | mgt01 |  |
| Hive Gateway | mgt01, dn[01-04] |  |
| Hive Metastore Server | mgt01 |  |
| Hive WebHCat | mgt01 |  |
| Hue | mgt01 |  |
| Activity Monitor | mgt01 |  |

Table 4.2 Assignment of roles in the cluster

The result shown in figure 4.8

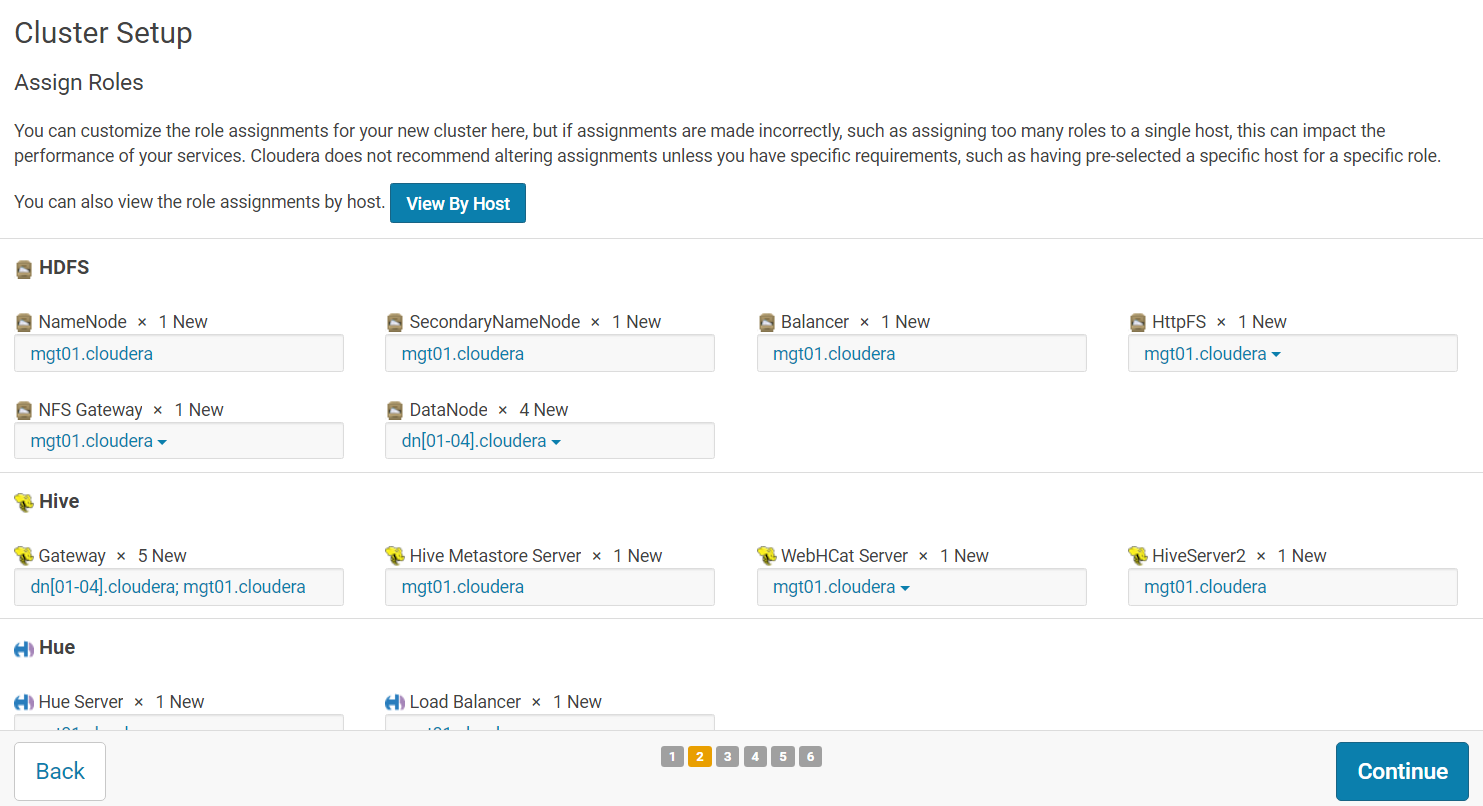


Figure 4.8 Assign Role for nodes in cluster

1. Database Setup: to configure and test the database connection for each service. Here we use an embedded PostgreSQL database. And the important point here to take a note about the credentials the setting by default as we may need it for later use.

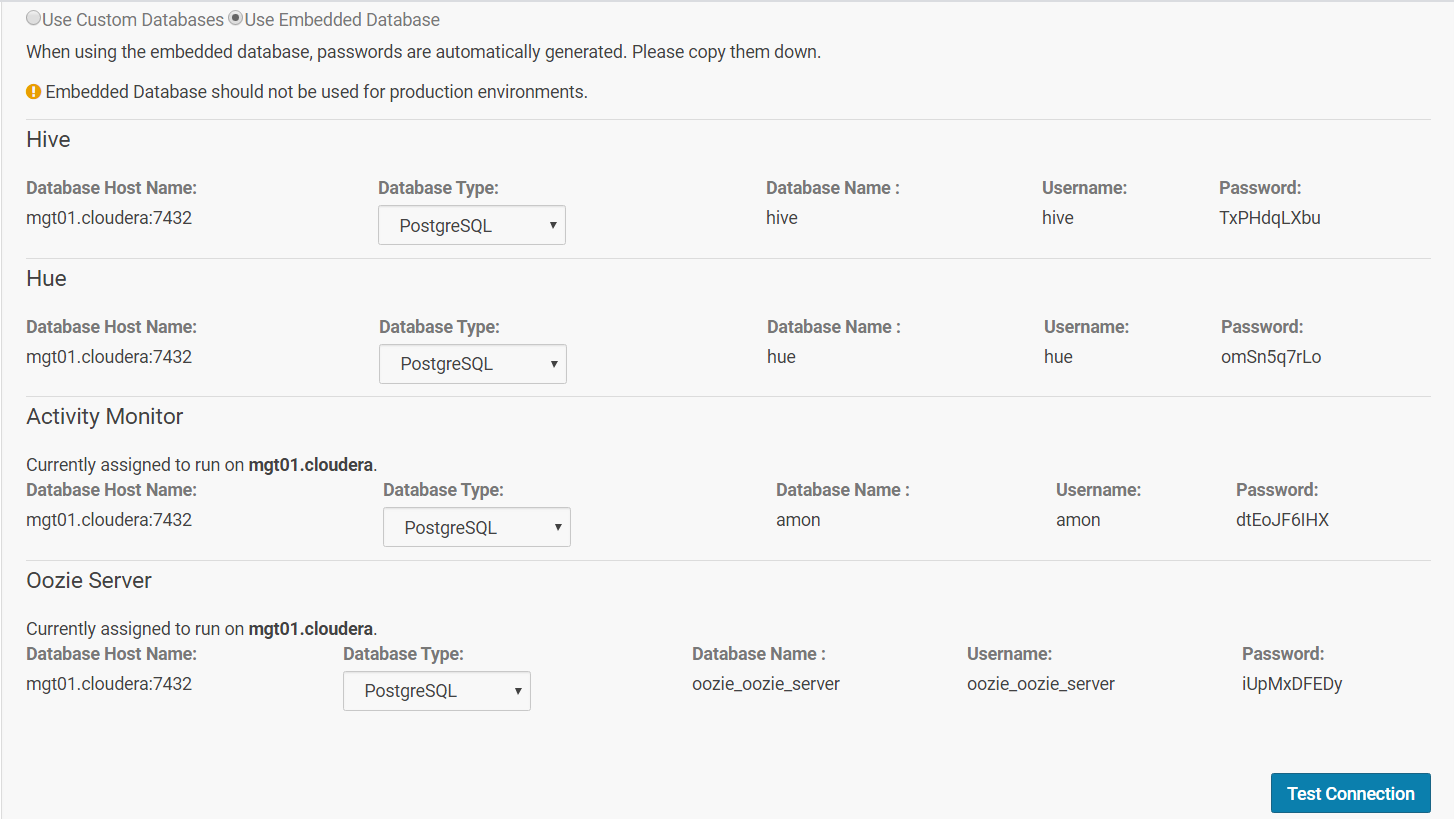


Figure 4.9 The default setting for embedded database

1. We can change the block size and directory for services. But we left every thing as default.

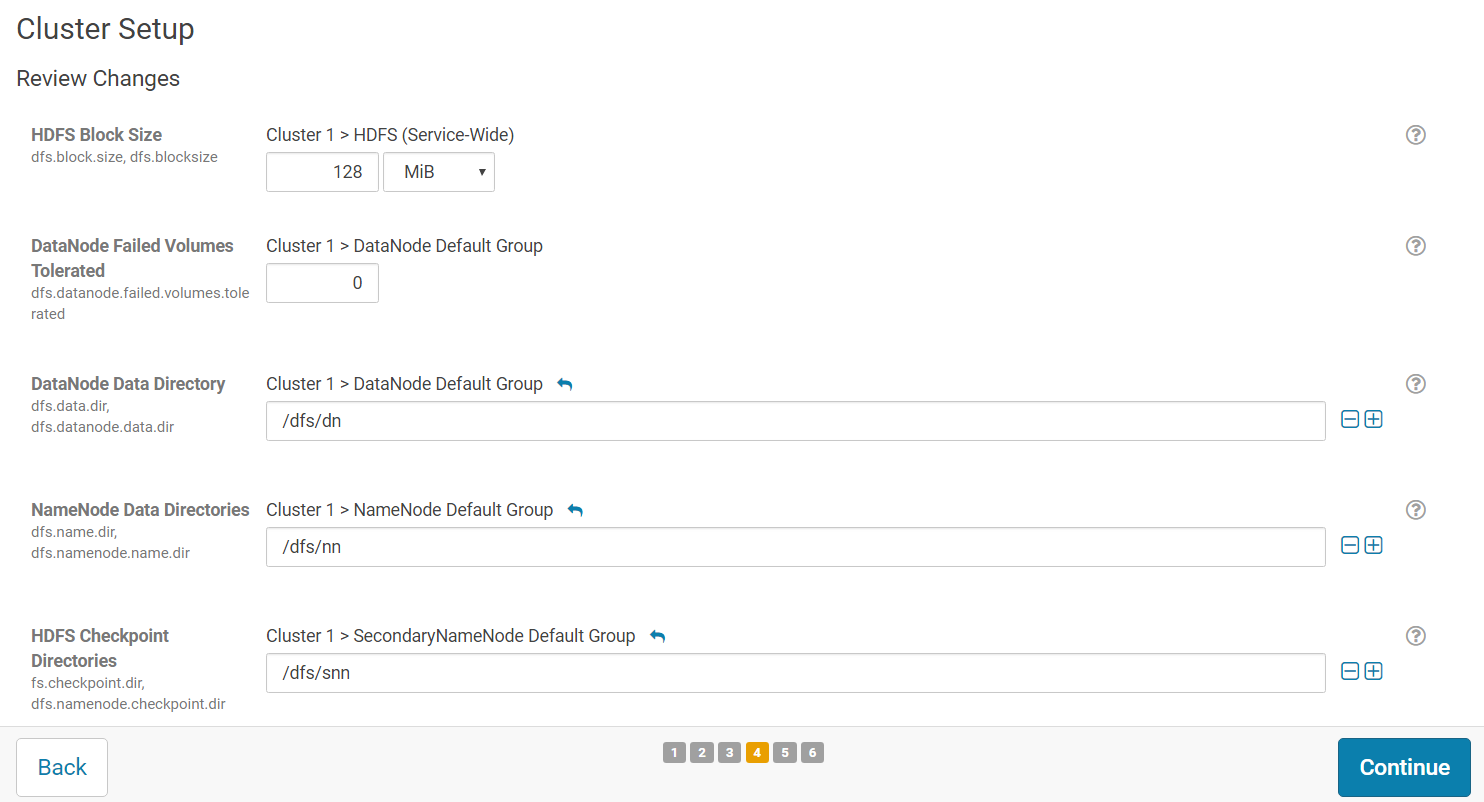


Figure 4.10 Setting the block size

1. Cloudera Manager(CM) will run all services for first time

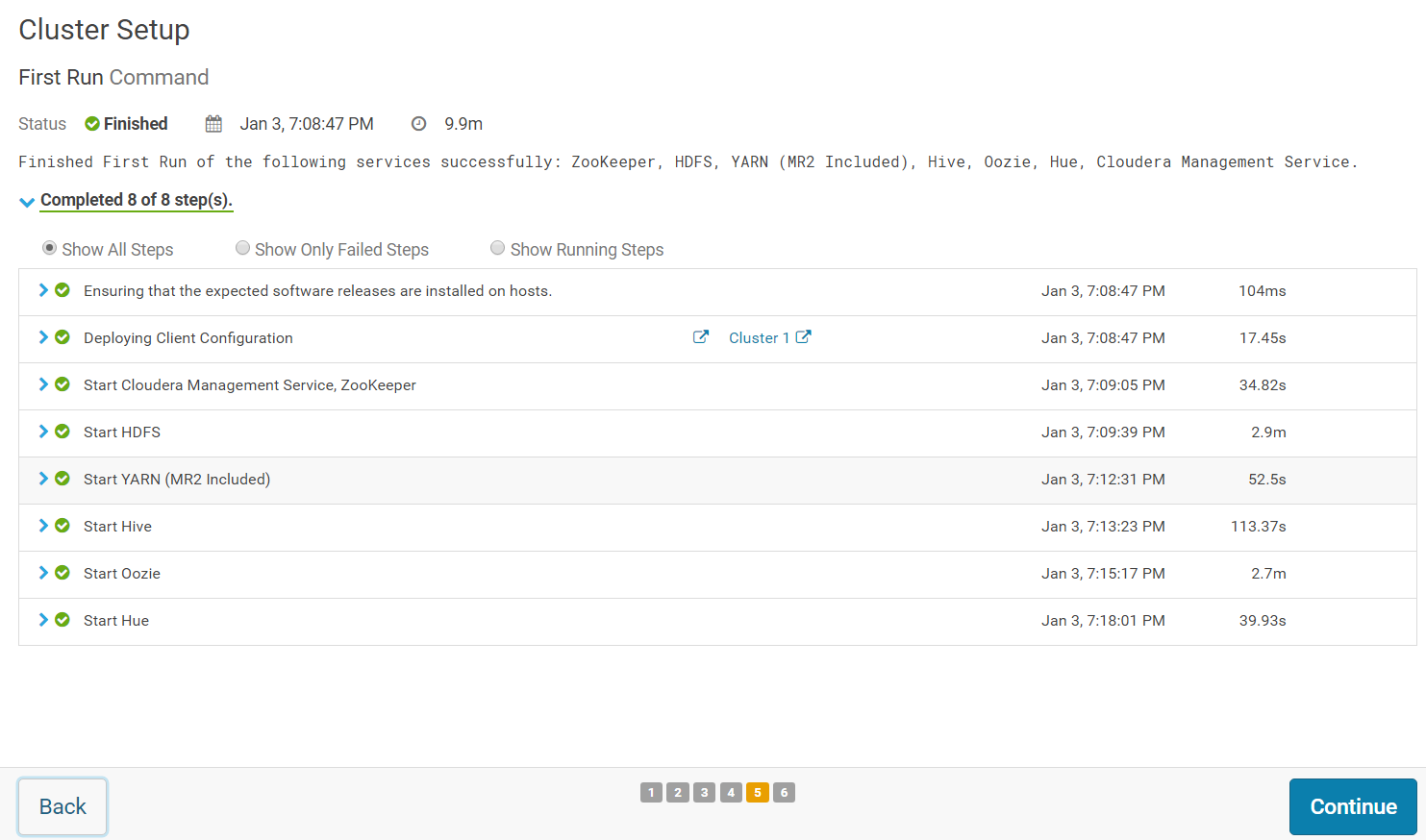


Figure 4.11 CM run all services to check

1. Then it will take us to the main administrative screen of Cloudera manager wich give us a full overview of our cluster. Where we can start, stop, restart, install, and delete our services.

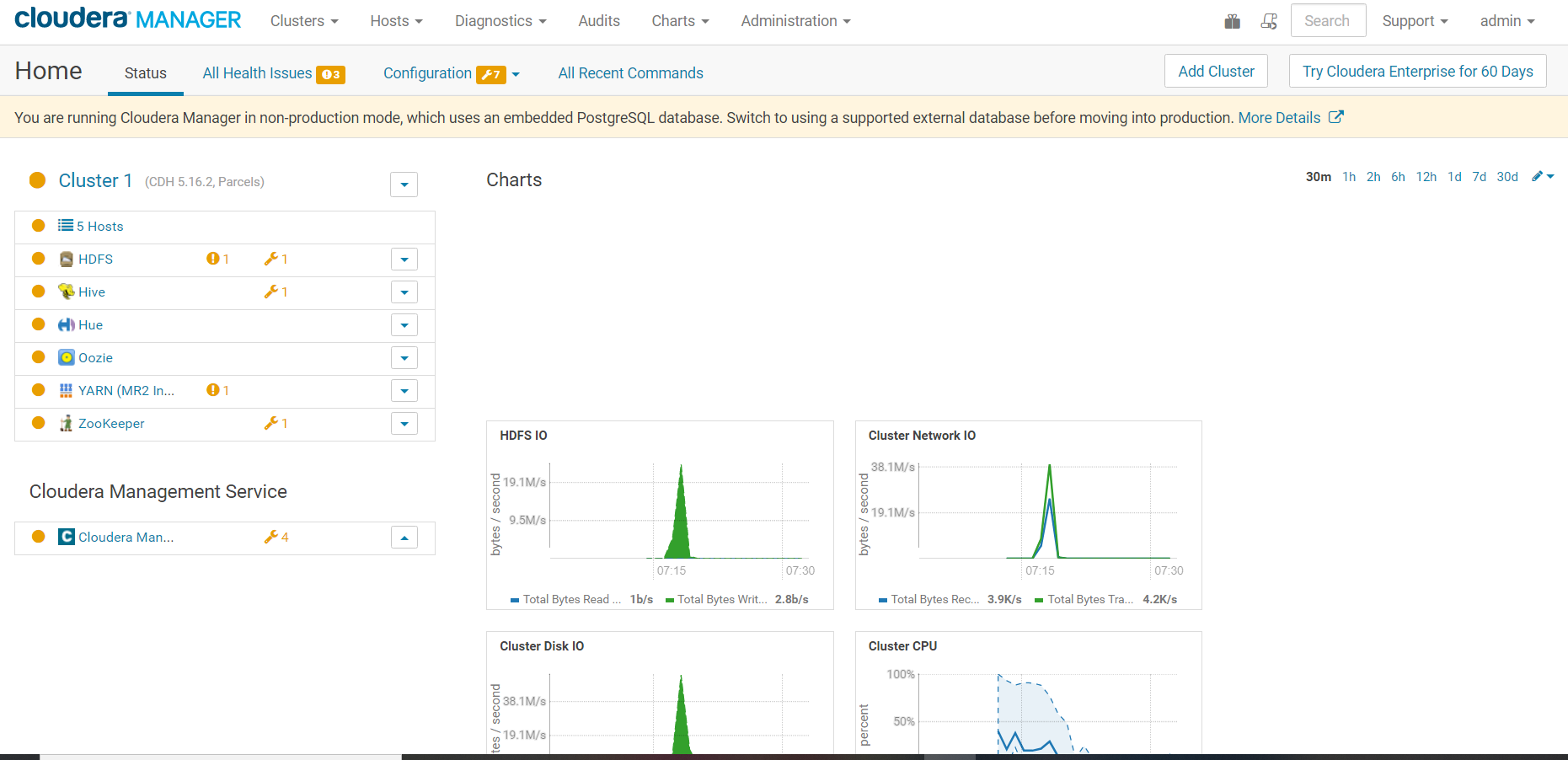


Figure 4.12 Cloudera manager main screen

**Installing NiFi in CDH**

1. **Get the CFM CSD files**

First we have to connect to Cloudera Manager node using the root account, we had adding the following setting:

cd /opt/cloudera/csd

wget http://archive.cloudera.com/CFM/csd/1.0.0.0/NIFI-1.9.0.1.0.0.0-90.jar

wget http://archive.cloudera.com/CFM/csd/1.0.0.0/NIFICA-1.9.0.1.0.0.0-90.jar

wget http://archive.cloudera.com/CFM/csd/1.0.0.0/NIFIREGISTRY-0.3.0.1.0.0.0-90.jar

chown cloudera-scm:cloudera-scm NIFI\*.jar

chmod 644 NIFI\*.jar

service cloudera-scm-server restart

when the restart operation done, we loged to CM and restart the Cloudera Management Service:

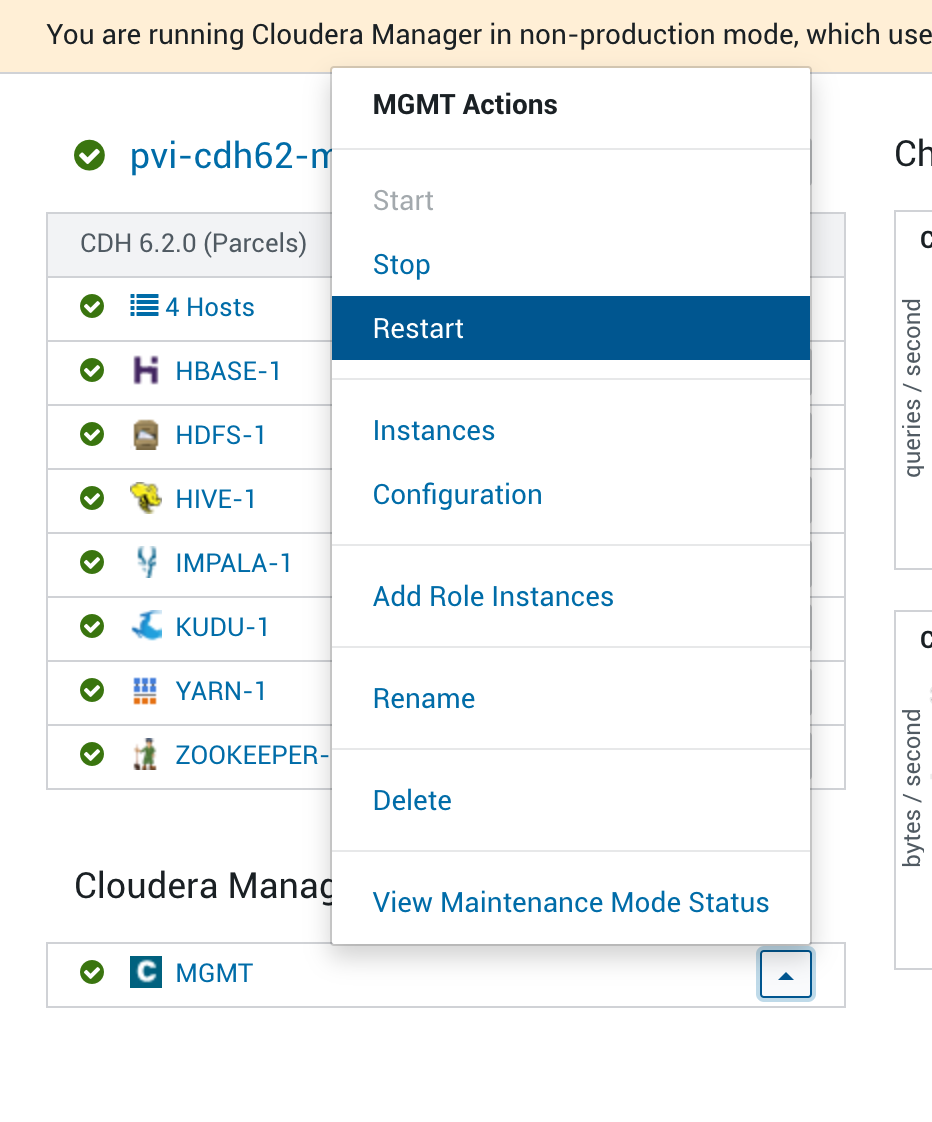


Figure 4.13 Restarting the Cloudera Management Service

1. **Get the CFM Parcel**

Getting CFM parcel is important to install NiFi on ClouderaCDH, we achieved that from parcels configuration (parcel icon top right):

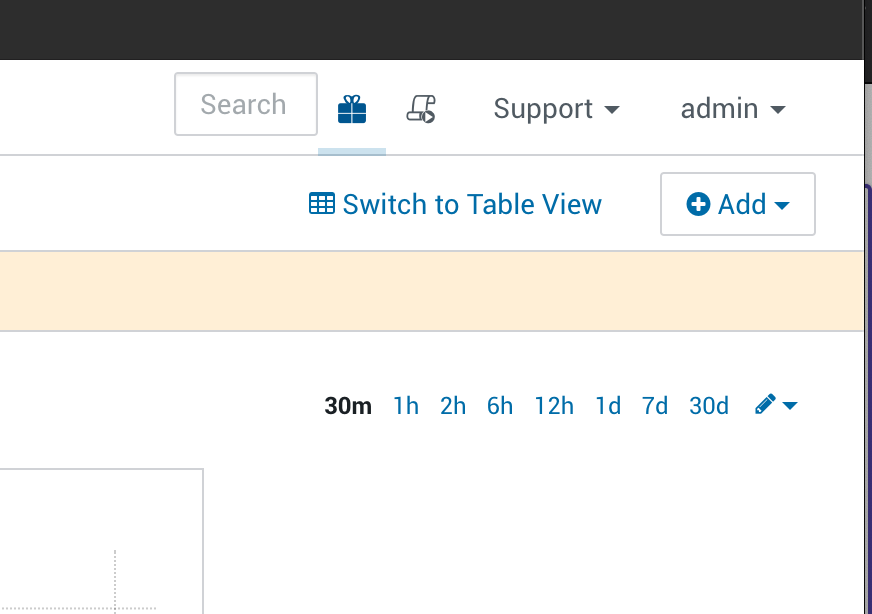


Figure 4.14 adding CFM parcel

NiFi parcel configuration obtained <http://archive.cloudera.com/CFM/parcels/1.0.0.0/> and added to the parcel repository list. click on the “+” icon:



Figure 4.15 CFM parcel URL

By clicking “Check for New Parcels” button, we should see a CFM parcel, then we manage to click “Download”. After the download completes, it will distribute and activate the parcel.

1. **Install Services**

We need to add NiFi service, so from CM main page, and “Add Service”:

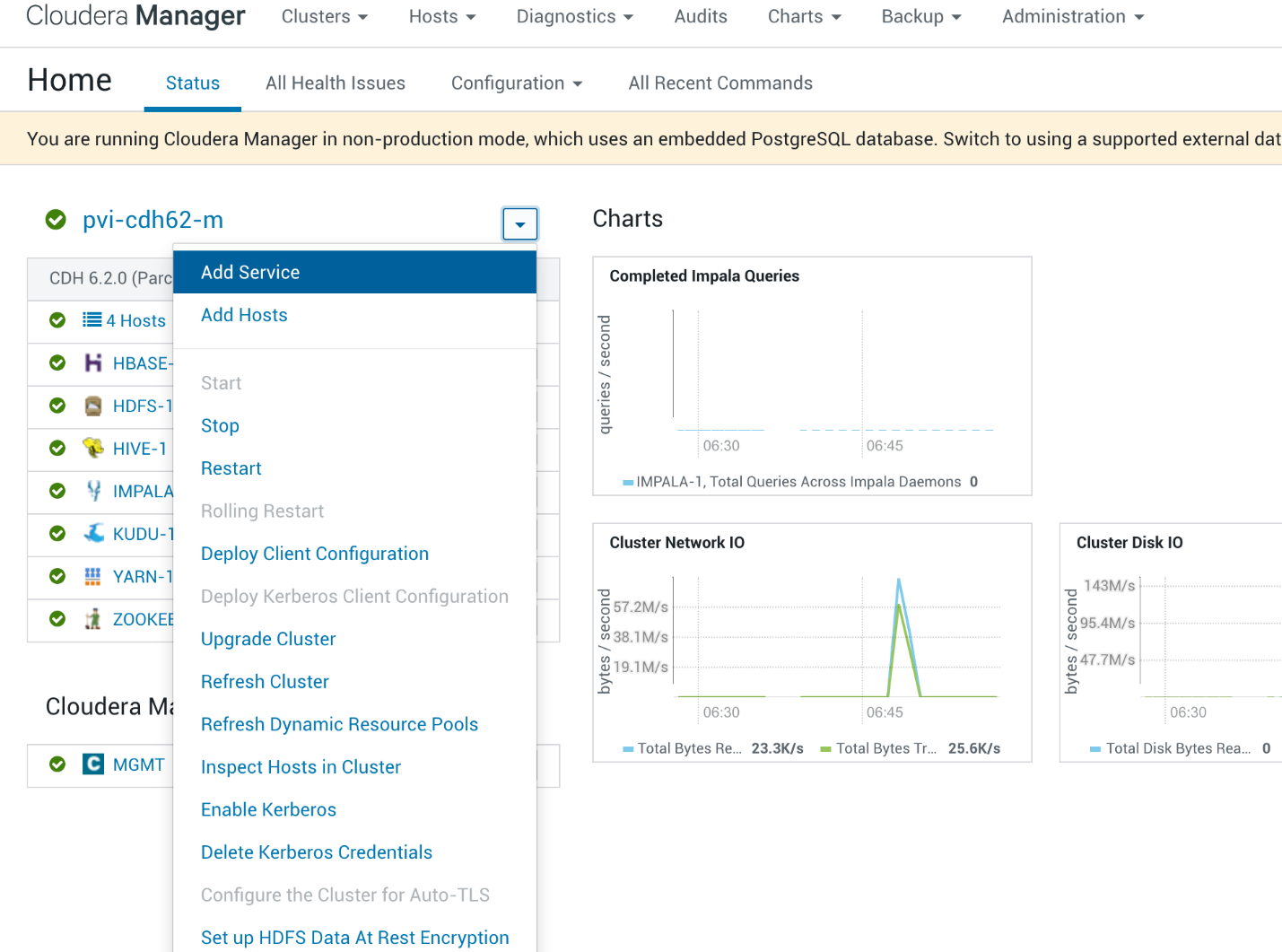


Figure 4.16 Adding NiFi service

Then we select the Nifi service, and specify the host onto which we want to install Nifi:

1. **Profit**

If we want to see main screen we had to go to Nifi URL localhost:9090 :

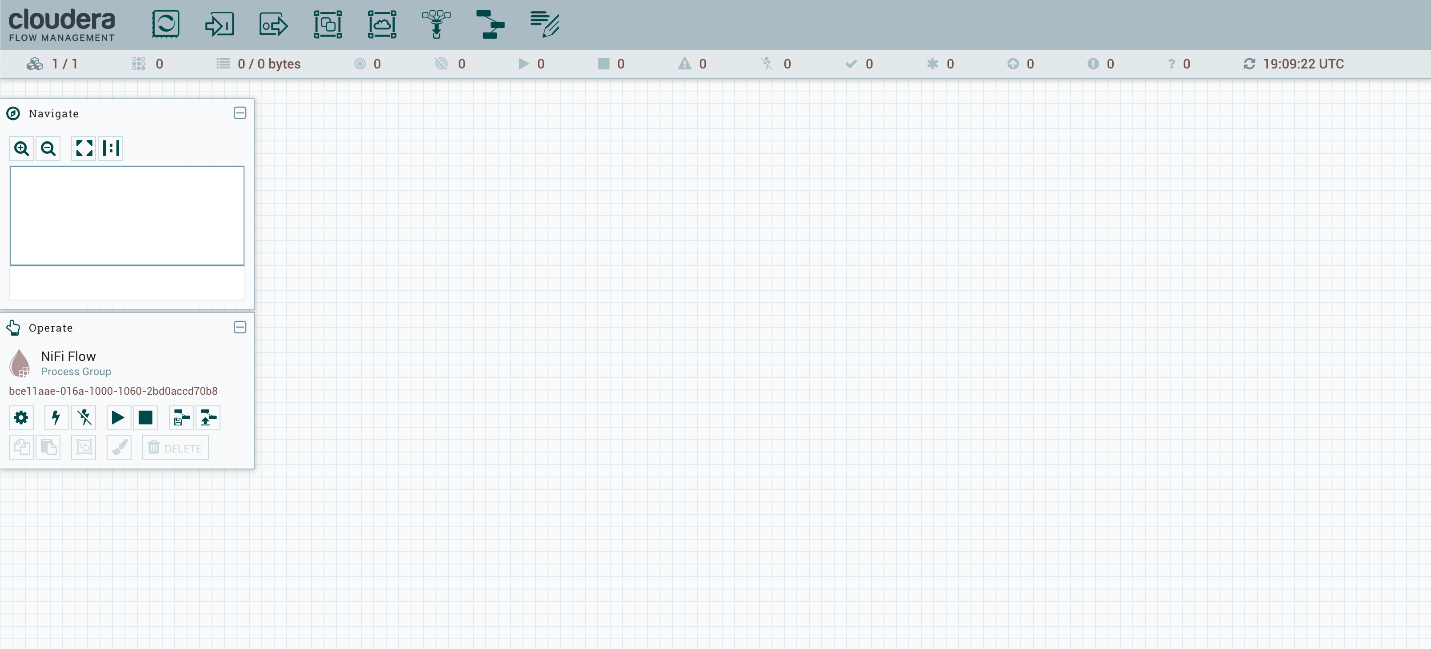


Figure 4.17 Apache NiFi main screen

The property nifi.cluster.flow.election.max.wait.time had been setting to 5 minutes by default, which can be confusing by giving rise to a log message like the cluster is still in the process of voting. to modify this property we can configure the nifi.properties.xml in CM, as follows:

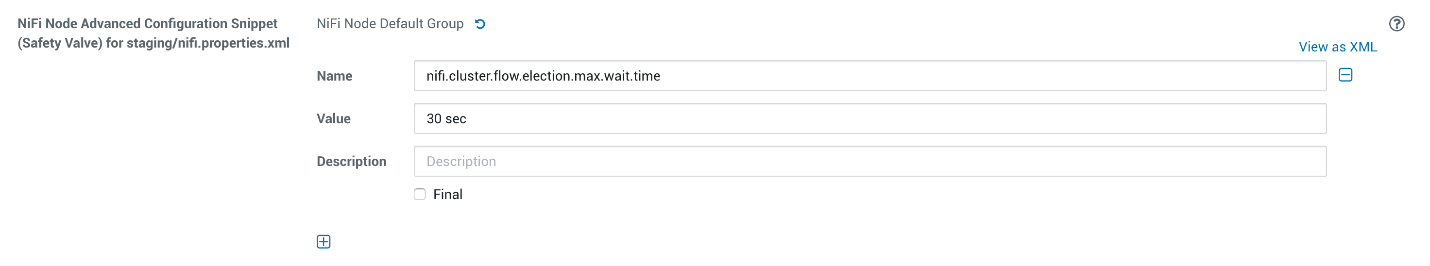


Figure 4.18 Setting NiFi Cluster property

At last we have to restart NiFi Service to let configuration take affect.

# 5. System Implementation

This chapter discuss how we implement all those systems and configure them to provide an integrated platform that will be effective in the three operations that we mentioned before: integration data, storing data, analyze and visualization.

So we used apache NiFi for ingestion of the data, Apache Hadoop and Apache Hive to the storage system, and Apache Spark for analyzing and visualization. Also, we install some of Hadoop echo system in order of building and utilizing the platform: Oozei, Zookeeper, Ambari, HUE. We install all of that on Cloudera CDH through Cloudera Manager.

## 5.1 Implementing Data Ingestion Layer

As we said before, we installed Apache NiFi on Cloudera CDH by adding CFM to it, and this allows us to unify entire subsystems into one system because Cloudera CDH without adding CFM will allow us to store, analyze, visualize and even ingestion but not with NiFi. The alternative was to install NiFi on other devices and connect to the main node after configuring the channel. Moreover, taking a round trip which is the bottleneck of speed, which means intensive use for network resources, and that will slow down the system.

In the Ingestion process, we ingest data from three different sources:

* the CSV file, which contains a different file from many offices and departments
* MySQL database contains important information about crops and lands.
* Weather data from Openweathermap.com using API.

### 5.1.1 Developing Process to Ingest CSV files

Processors that ingest data from the CSV data source to the main repository in Apache Hive are all placed into a single process group. A process group enables the grouping of processors into one. These processors extract data from each CSV file and stream it directly into Apache Hive.

The following steps below are followed to build the data ingestion platform for the CSV data source.

* + - 1. The GetFile processor is added and configured to ingest data from the external location of the PAE data to Apache Nifi. it shown in figure 5.1.

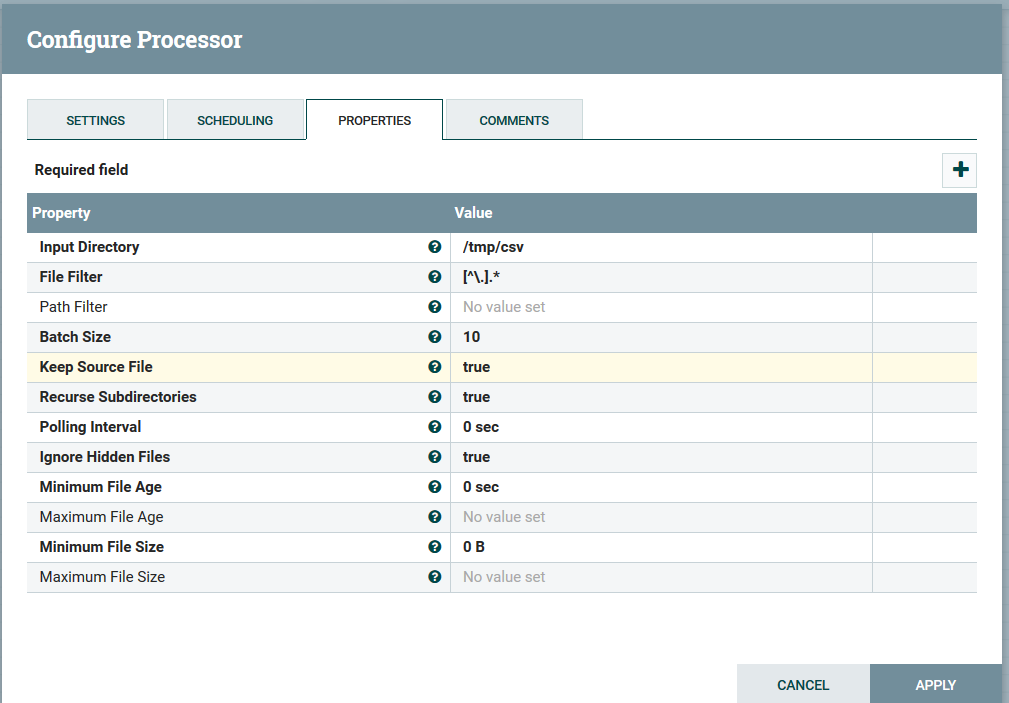


Figure 5.1 GetFile processor to get CSV data to Apache NiFi

The Input Directory: specifies the location of the CSV file containing the PAE data. It is located in /tmp/csv directory.

* + - 1. The InferAvroSchema is added and configured to examine the content of the file before it is ingested to the main repository in Hive.

Figure 5.2 showing the InferAvroSchema for Cotton Harvest Area with configurations.

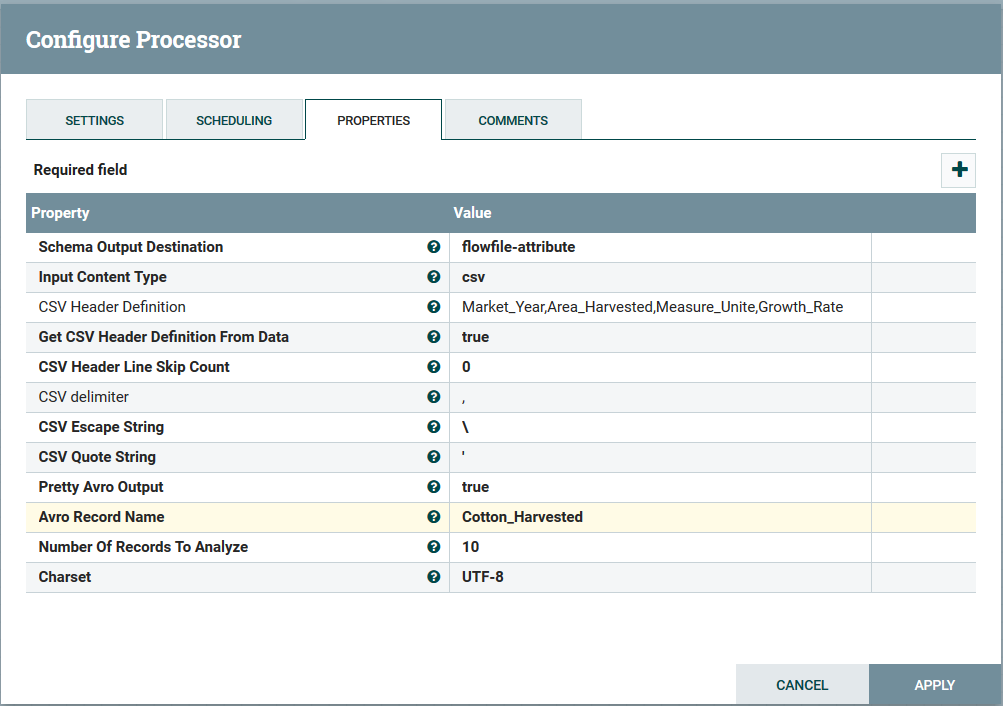


Figure 5.2 InferAcroSchema processor to examine the content of the file

The **Schema Output Destination** specifies whether the Avro schema is written as a new attribute “inferred.avro.schema.”

The **Input Content-Type** specifies the content type of the incoming flow file. It supports files that are either in CSV or JSON.

The **CSV Header Definition** specifies a list of columns separated by commas that serve as the Header for the CSV file. This applies only to CSV files. These columns will be retrieved as the header definition of the CSV file.

* + - 1. The ConvertCSVToAvro processor is configured to convert the CSV file to Avro format. This is because the PutHiveStreaming processor only understands data in Avro format.

the ConvertCSVToAvro processor and its configuration shown on figure 5.3:

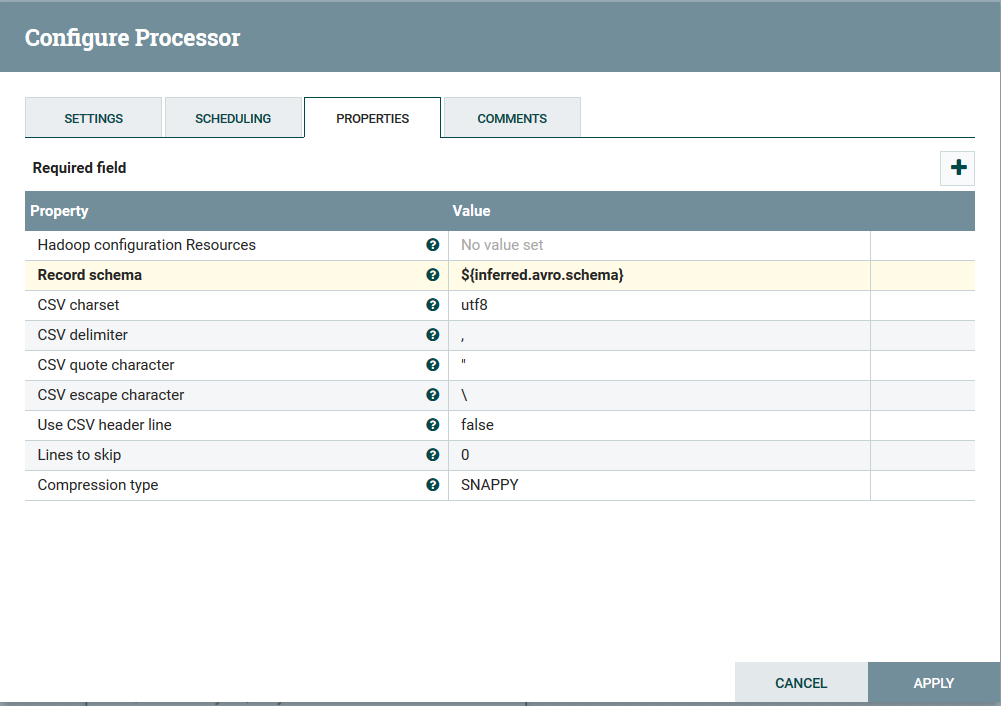


Figure 5.3 ConvertCSVToAvro processor to convert the CSV file to AVRO format

The **Hadoop configuration Resources** specifies the list of files that contain Hadoop’s configuration. It will use the default configuration to search for the files when left empty.

The **Record schema** specifies the outgoing Avro schema for each record created from a CSV row.

* + - 1. The PutHiveStreaming processor is configured to ingest the data to Hive. the PutHiveStreaming processor with configurations ollustrated in figure 5.4



Figure 5.4 PutHiveStreaming processor to Ingest PAE data to Hive

The following code snippet is the SQL syntax, which creates the table to receive CSV input file in Hive.

***Create table har\_cotton(Market\_Year int, Area\_Harvested Float, Measure\_Unite String, Growth\_Rate Float) CLUSTERED BY(Market\_Year) INTO 2 BUCKETS STORED AS ORC tblproperties("transactional"="true");***

* + - 1. The LogAttribute processor is configured to tell whether the data is successfully ingested or not.

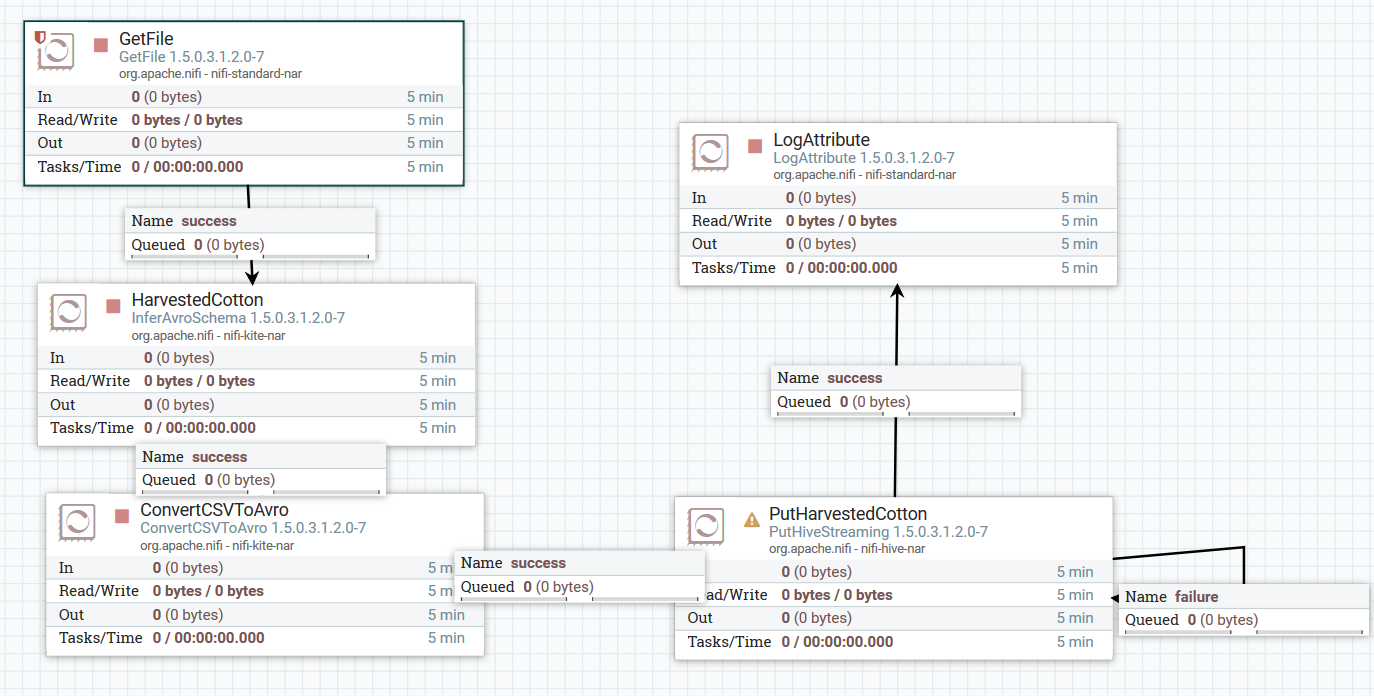


Figure 5.5 a complete flow with processors to ingest CSV data to the main repository

we did this for a cotton harvested area by year, but of course, the table 5.1 show the full list for tables needed to ingest csv file

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| CSV file by year | cotton | Sorghum | Millet | Wheat |
| Area Harvested | har\_cotton | har\_sorg | har\_ millet | har\_wheat |
| Beginning Stock | bstock\_cotton | bstock\_ sorg | bstock\_ millet | bstock\_ wheat |
| Ending Stock | estock\_cotton | estock\_ sorg | estock\_ millet | estock\_ wheat |
| Exports | exports\_cotton | exports\_ sorg | exports\_ millet | exports\_ wheat |
| Imports | imports\_cotton | imports\_ sorg | imports\_ millet | imports\_ wheat |
| Production | production\_cotton | production\_ sorg | production\_ millet | production\_ wheat |
| Yield | yield\_cotton | yield\_ sorg | yield\_ millet | yield\_ wheat |

Table 5.1 The database tables name for all CSV files

### 5.1.2 Developing Process to Ingest MySQL to Hive

## 5.1 Introduction

Family planning data is selected and anayzed as a case study to demonstrate the importance of such platform in the ministry of health and sanitation.

Family Planning methods are birth control measures to prevent pregnancy for both men and women. Also known as contraceptive methods, it helps both men and women to determine the number and spacing of children. Family planning can also help to delay pregnancies in young women that have increased risk of health problems. There are differently available family planning or contraceptive methods which help both men and women to plan their family and prevent unwanted pregnancy. Family planning method is one of the 10th great public health achievements of the 20th century.

According to the World Health Organization (WHO) (2018), the use of Family Planning/contraceptive methods to prevent pregnancy has led to a series of benefits.

* Family Planning reduces infant mortality by preventing births that are not timely and closely spaced which has been one of the factors of infant mortality.[36]
* Because family planning reduces the risk of unwanted pregnancy among women living with HIV, it has led to situations where the number of infected babies or orphans is reduced. The use of Condoms does not only protect against unwanted pregnancy but also against sexually transmitted disease (STDs) like HIV, syphilis, herpes, etc.[36]
* Family planning methods have enabled young women to pursue education and gain employment in society. With small families, parents can properly take care of their children.[36]
* Family planning has contributed to slow the growth of the population in the world. With this, it helps countries to plan well according to resources available to take care of their population.[36]

## 5.2 Common Family Planning Methods in Sierra Leone

In Sierra Leone, different family planning services are provided to help prevent pregnancy. Family planning services are offered at no cost to patients. The government and other donors sponsor these services. They are considered to be under the free health care program implemented by the government of Sierra Leone, and it’s donors.

The analysis will consider the following family planning methods in Sierra Leone.

* **Oral contraceptive Pills:** Oral Contraceptive Pills is one of the family planning methods that have to do with hormones. Oral Contraceptive Pills are distributed to women, and a single pill is taken daily to prevent pregnancy. Oral Contraceptive Pills are more effective when taken at the same time every day and when no pill is missed. Normally a pack of thirty pills which lasts for the entire month is given to women.

Oral Contraceptive Pills is of the following types:

Combined Oral Contraceptive Pills (COCs): This type of oral contraceptive pills combine two synthetic hormones which are estrogen and progestin. This pill with combined hormones is almost 100% effective if taken correctly.

Progestin-Only Pills: As its name implies, this pill contains only one synthetic hormone known as progestin. Progestin-only pills are more suitable for women who are breastfeeding because they are more effective for them and will not reduce the production of breast milk. However, they are not very effective compared to the progestin-only pills for women who are not breastfeeding.

**Injectables:** For Injectable contraceptive method, women are given an injection on the arm or buttock once every 1,2 or 3 months. Injectables are also said to be close to 100% effective when women come back on-time to take another injection.[38]

**Implants:** In the implants contraceptive method, small flexible rods or capsules are inserted under the skin of the upper arm of a woman. The inserted implants normally provide protection for three to five years. After the implant is inserted, no further cost or additional protection is needed. After the period has expired, the old implant is removed, and a new one is inserted if the woman still needs the contraceptive protection. A woman becomes unprotected and fertile again for pregnancy almost immediately the implants are removed.[38]

**Male Condoms:** Male Condoms is a barrier contraceptive method that physically prevents sperm and egg from meeting. A Male condom is a sheath or covering that fit over a man’s penis. This method is about 98% effective if the condom is used correctly. It is also suitable for protecting against sexually transmitted diseases (STD)

**Female Condoms:** A female condom is a contraception method that stops the sperm that comes out of a man’s penis from entering a woman’s vagina when he ejaculates. The sperm goes into the female condom which is a loose-fitting sheath that fits inside a woman’s vagina. It can also help in protecting against sexually transmitted diseases (STDs) such as HIV, syphilis, etc. If used correctly, it provides 90% protection against unwanted pregnancy. [38]

**Intrauterine Device (IUD):** IUDs are small T-shaped plastic devices that are inserted into the uterus of a woman and steadily releases small amounts of levonorgestrel each day. The most common IUDs contain copper and thicken cervical mucous to prevent sperm and egg from meeting.IUDs can provide protection for five to twelve years. [38]

## 5.3 Relationship on family planning methods among the three data sources.

All three of the systems captures information about planning services/methods. This shows the extent to which all of them are related to each other and the need to store data about all of them in one location for efficient and effective analysis is very important. This enables ministry officials to have a clear picture of what is happening in all facilities of the entire ministry of health and sanitation.

The District Health Information System (DHIS2) contains information about the number of patients who went for family planning services within a specific period. Information about the facilities where the patients went to is also captured.

The Logistics Management Information System (LMIS) contains information about family planning health commodities administered to patients within a specific period. It contains information like the name of the product, the quantity administered to patients, the quantity that remains in the health facility within a specific period. The LMIS also captures information about the facility which administered the family planning health commodity.

The Integrated Human Resource Information System (IHRIS) contains information about health staff who are experienced in providing family planning services. It captures information about the qualification of the health staff including the educational qualification, work experience, various training attended, etc.

## 5.4 Data Selection

Family planning data related to each of the data sources is selected and saved to the main repository. The process of selecting family planning related data for each of the data sources from the main repository involves querying the data in Hive query language (i.e., HiveQL) and saving the result of the query also to the main repository in Apache Hive. The results for each of the three data sources are saved on different tables in Hive. At the end of the selection process, three different tables (i.e., one for each data source) are derived.

The following steps are followed to select and save family planning related data for each of the three data sources:

**District Health Information System2 (DHIS2):**

1. With HiveQL, create a table which will store family planning related records selected from different tables of the DHIS2 data source in the main repository.

***CREATE TABLE IF NOT EXISTS datavalue1***

***(organisationunitid int,name varchar(255),idlevel2 int,idlevel3 int,idlevel4 int,dataelementid int,categoryoptioncomboid int,periodid int,periodtypeid int,level int,value double)***

***ROW FORMAT DELIMITED FIELDS TERMINATED BY ',' STORED AS ORC;***

The HIVEQL query above creates a table called datavalue1.The table is converted to ORC format to ensure that HIVE understands the data in it.Optimized Row Format(ORC) ensures that data is stored efficiently in Hive.

1. Select records related to family planning from different tables of the DHIS2 data source and insert to the dhis\_datavalue1 table.

***INSERT INTO TABLE datavalue1***

***SELECT***

***dhis\_aggregated\_datavalue.organisationunitid,***

***dhis\_organisationunit\_struct.name,***

***dhis\_orgunit\_structure.idlevel2,***

***dhis\_orgunit\_structure.idlevel3,***

***dhis\_orgunit\_structure.idlevel4,***

***dhis\_aggregated\_datavalue.dataelementid,***

***dhis\_aggregated\_datavalue.categoryoptioncomboid,***

***dhis\_aggregated\_datavalue.periodid,***

***dhis\_aggregated\_datavalue.periodtypeid,***

***dhis\_aggregated\_datavalue.level,***

***dhis\_aggregated\_datavalue.value***

***FROM***

***dhis\_aggregated\_datavalue,***

***dhis\_orgunit\_structure***

***WHERE***

***dhis\_organisationunit\_struct.organisationunitid = dhis\_aggregated\_datavalue.organisationunitid***

***AND dhis\_aggregated\_datavalue.dataelementid=1748***

***AND dhis\_aggregated\_datavalue.categoryoptioncomboid BETWEEN 1631 AND 1651***

***AND dhis\_aggregated\_datavalue.periodtypeid IN (3,4,5,6);***

The HiveQL syntax above selects data from multiple tables of the DHIS2 data source and inserts them into the dhis\_datavalue1 table.

Steps 1 and 2 are repeated to create a final table that contains all the fields with family planning related data for the DHIS2.

Below are the other Hive queries.

***CREATE TABLE IF NOT EXISTS dhis\_datavalue(organisationunitid int,orgunitname varchar(255),idlevel2 int,idlevel3 int,idlevel4 int,dataelementid int,dataelementname varchar(255),categoryoptioncomboid int,periodid int,startdate date,enddate date,periodname varchar(100),periodtypeid int,periodtypename varchar(100),level int,value double)ROW FORMAT DELIMITED FIELDS TERMINATED BY ',' STORED AS ORC;***

The HiveQL query above creates a table called dhis\_datavalue and converts it to ORC format.

***INSERT INTO TABLE dhis\_datavalue***

***SELECT***

***datavalueall.organisationunitid,***

***datavalueall.name AS orgunitname,***

***datavalueall.idlevel2,***

***datavalueall.idlevel3,***

***datavalueall.idlevel4,***

***datavalueall.dataelementid,***

***dhis\_dataelement.name AS dataelementname,***

***datavalueall.categoryoptioncomboid,***

***datavalueall.periodid,***

***dhis\_period.startdate,***

***dhis\_period.enddate,***

***dhis\_periodtype.name AS periodname,***

***dhis\_datavalueall.periodtypeid,***

***dhis\_periodtype.name AS periodtypename,***

***datavalueall.level,***

***datavalueall.value***

***FROM***

***datavalueall,***

***dhis\_period,***

***dhis\_periodtype,***

***dhis\_dataelement***

***WHERE***

***datavalueall.dataelementid = dhis\_dataelement.dataelementid AND***

***datavalueall.periodtypeid = dhis\_periodtype.periodtypeid AND***

***datavalueall.periodid = dhis\_period.periodid***

***AND dhis\_period.startdate between '2012-07-01' AND '2014-12-01'***

***AND dhis\_period.enddate between '2012-07-31' AND '2014-12-31'***

***ORDER BY dhis\_period.startdate;***

From the HiveQL query above, it selects data from multiple tables which contains family planning related data of the DHIS2 data source and inserts them to dhis\_datavalue table.

**Integrated Human Resource Information System (IHRIS):**

* With HIVE query language, create a table and insert family planning related data of the IHRIS data source to it.

***CREATE TABLE IF NOT EXISTS hris\_data***

***(district varchar(100),facility varchar(100),personid varchar(50),firstname varchar(100),othername varchar(100),surname varchar(100),title varchar(100))***

***ROW FORMAT DELIMITED FIELDS TERMINATED BY ',' STORED AS ORC;***

The HiveQL query above creates a table called hris\_data.The table is stored in ORC format to ensure that Apache Hive easily understands the data.

* Select family planning records from different tables of the IHRIS data source and insert them to the hris\_data table.

***INSERT INTO TABLE hris\_hrisdata***

***SELECT (hris\_district.name) district, (hris\_facility.name) facility, (hris\_healthstaff.id) personid, hris\_healthstaff.firstname,hris\_healthstaff.othername,hris\_healthstaff.surname,hris\_staff\_position.title***

***FROM hris\_district,hris\_facility,hris\_staff\_position***

***WHERE***

***hris\_facility.location=hris\_district.id***

***AND hris\_staff\_position.facility=hris\_facility.id***

***AND hris\_staff\_position.parent=hris\_healthstaff.id***

***AND hris\_staff\_position=hris\_position.id;***

The Hive Query Language above selects data from multiple tables of the IHRIS data source and insert them to the hris\_hrisdata table.

Logistics Management Information System (LMIS):

* Create a table to store family planning related records for the LMIS data source.

***CREATE TABLE IF NOT EXISTS lmisdata***

***(fromdate date,todate date,facility varchar(150),district varchar(100),product varchar(100),openingbal int,quantityrec int,lossadj int,quadisp int,closingbal int)***

***ROW FORMAT DELIMITED FIELDS TERMINATED BY ',' STORED AS ORC;***

The Hive Query Language above creates a table with appropriate fields to store family planning data for the LMIS data source.

* Select family planning related records from the “lmis” table and insert to the lmisdata table.

***INSERT INTO TABLE lmisdata***

***SELECT fromdate,todate,facility,district,product,openingbal,quantityrec,lossadj,quadisp,***

***closingbal***

***FROM lmis***

***WHERE product IN ('Jadelle Implant/Norplant','Female Condoms','Male Condoms','Depo Provera','Microgynon Pills','Microlut Pills');***

## 5.5 Analysis and Visualization

### 5.5.1 Goal of the analysis

This analysis aims to investigate the level of inconsistency of family planning data in both the District Health Information System2 (DHIS2) and the Logistics Management Information System (LMIS) data sources. This process involves selecting data from two different data sources and analyzing it to determine the level of inconsistency. All of these processes are executed in one platform. The analysis also goes further to investigate the possible causes leading to the inconsistency on family planning data from both data sources. The investigation is done for a period of four months (i.e., January 2014 to April 2014). Two districts (i.e. Pujehun and Tonkolili districts) are considered during the analysis. The analysis covers the following family planning methods in Sierra Leone:

* Oral Contraceptive Pills
* Implants
* Injectable
* Male Condoms
* Female Condoms

The analysis involves looking at the number of patients who received family planning services from health facilities and compare that to the quantity of family planning health commodities administered by health facilities to patients. The number of patients who received family planning services from health facilities is captured in the DHIS2 while the quantity of family planning health commodities administered is captured in the LMIS.

The table below shows family planning health commodity administered for each family planning service.

|  |  |
| --- | --- |
| **Family Planning Method/Service(DHIS)** | **Health commodity for each service (LMIS)** |
| Oral contraceptive | Microgynon Pills or Microlut Pills |
| Implants | Jadelle Implant |
| Injectable | Depo Provera |
| Male Condoms | Male Condoms |
| Female Condom | Female Condoms |

Table 5. 1 Family planning services and health commodities

To determine whether the number of patients who received family planning services(i.e cases) and the quatity of family planning health commodities administered(i.e dispensed) are consistent, the following conditions must be met.

* The number of family planning cases must not be greater than the quantity of family planning health commodities administered (i.e. Consumption). The number of family planning cases must be less than or equal to the quantity of family planning health commodities administered.
* If the first condition is satisfied, the percentage difference between the number of family planning cases and the quantity of family planning health commodities administered must not be very large. For this analysis, Ten percent (10%) is considered to be the maximum percentage difference.

However, the second condition is not considered for both Male and Female Condoms because the there is no specified fixed dosage of male and female condoms that should be administered to people.

The data is automatically considered to be inconsistent if the first condition is not satisfied.

This is illustrated in the flow chart below:

Is case >consumption

NO

YES

IS product=Male Condoms or Female Condoms

NO

YES

Is perdiff >10

Data is inconsistent

Data is consistent

Data is consistent

NO

YES

Data is inconsistent

Figure 5. 1 A flow chart illustrating conditions to determine whether the data is consistent or not

Oral Contraceptive: The number of people who visited the health facilities for Oral Contraceptive services should be less than or equal to the quantity of oral contraceptive drugs administered (i.e. Consumed).Either of the two drugs (i.e., Microgynon pills or Microlut Pills) is administered to patients. A single sachet which contains thirty (30) pills is administered to a patient for a whole month.

Implants: The number of patients who received implant family planning method/service must be equal to or slightly less than (i.e.10% difference) the quantity of Jadelle implant administered. A single Jadelle Implant is inserted into the skin of the upper arm of a woman.

Injectable: A single Depo Provera injection is administered on the arm or buttock of women who received injectable family planning method/services from health facilities. Therefore, the number of patients who received injectable family planning service must be equal to or slightly less than (i.e., 10% difference) the quantity of Injectable (i.e., Depo Provera) administered.

Male Condoms: There is no fixed dosage on the number of male condoms that should be distributed to patients, but to determine the level of consistency, the number of men who received male condoms family planning service must not be greater than the number of male condoms administered.

Female Condoms: For the female condom family planning method, there is also no fixed dosage on the number of female condoms to be administered to women. However, the number of female condoms administered must not be less than the number of women who received female condom family planning service.

### 5.5.2 Analysis to determine the level of inconsistency of family planning data

The analysis below compares the number of family planning cases from the DHIS2 data source with the quantity of family planning health commodities from the LMIS data source administered (i.e. consumed) by health facilities. This is done for both Pujehun and Tonkolili district. A negative difference indicates that there is inconsistency between the numbers of people who received family planning services (i.e., Cases) and the quantity of family planning health commodities administered (i.e., consumed) by health facilities. A positive difference and a percentage difference of less than or equal to 10% indicates that there is consistency. The periods under review are January 2014, February 2014, March 2014 and April 2014.

Apache SparkSQL queries are executed in Apache Zeppelin to analyze the data. The data is also visualized in Apache Zeppelin.

#### 5.5.2.1 Pujehun District

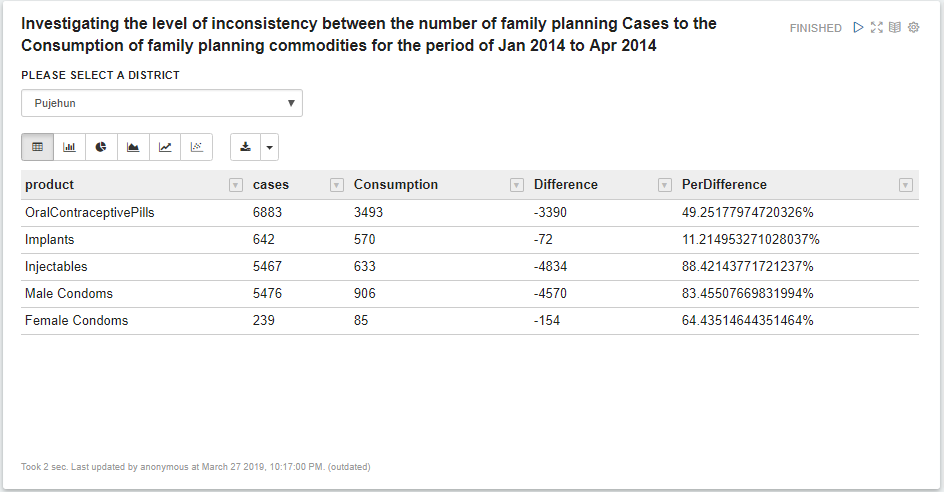


Table 5. 2 Cases, consumption, difference and percentage difference of family planning data for Pujehun district.

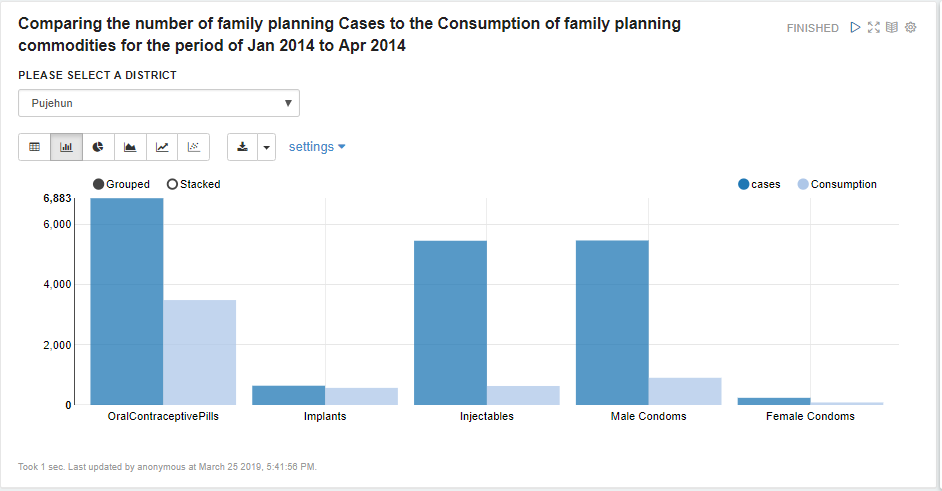


Figure 5. 2 a comparison between family planning cases and consumption of family planning health commodities for Pujehun district.

From the table and bar chart above, it clearly shows that the number of women who received Oral Contraceptive family planning service (6883) exceeds the quantity of Oral Contraceptive Pills consumed (3493).A negative difference of -3390 and a percentage difference of 49% clearly shows that there is a high level of inconsistency between the number of Oral Contraceptive Cases and the quantity of Oral Contraceptive pills administered (i.e., consumed).

The same applies for Implants with a negative difference of -72, Injectable with a negative difference of -4834, Male Condoms with a negative difference of -4570 and Finally Female Condoms with a negative difference of -154.

The analysis above concludes that there is a high level of inconsistency between the number of patients who received family planning services and the quantity of family planning health commodities administered by health facilities to those patients.

#### 5.5.2.2 Tonkolili District

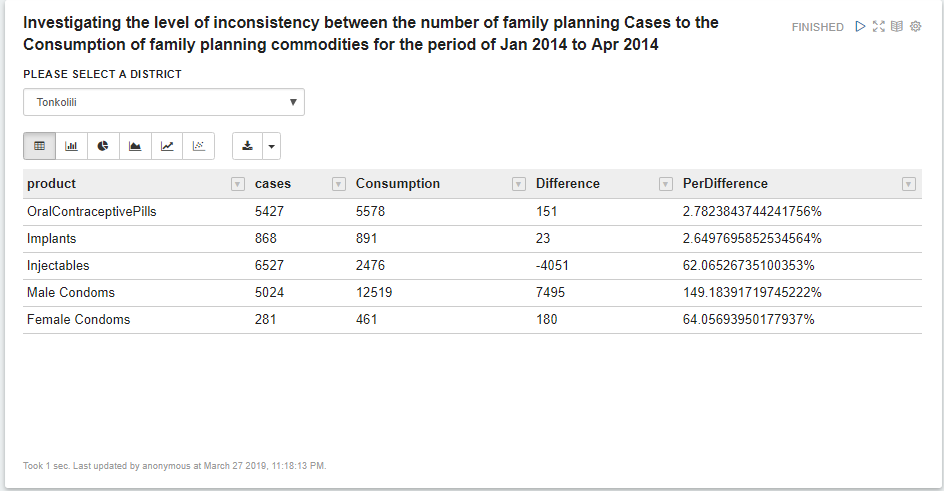


Table 5. 3 Cases, consumption, difference and percentage difference of family planning data for Tonkolili district.

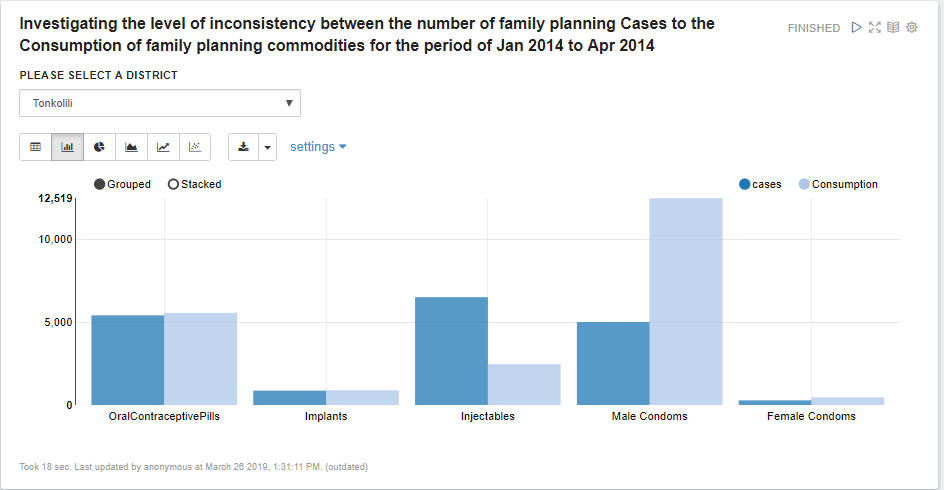


Figure 5. 3 a comparison between family planning cases and consumption of family planning health commodities for Tonkolili district.

The figures above compares the number of family planning cases and the quantity of family planning health commodities administered (i.e consumed) by health facilities from January 2014 to April 2014 in Tonkolili district.

The number of cases for Oral Contraceptive family planning service (5427) is less than the quantity of Oral Contraceptive Pills consumed (5578). This satisfies the criteria which state that the number of women who receive Oral Contraceptive Pills family planning services should not exceed the quantity of Oral Contraceptive Pills administered to them. A positive difference of 151 and a positive percentage difference of 2.78% indicates that the number of cases for Oral Contraceptive pills and the quantity of Oral Contraceptive Pills consumed are consistent.

The same applies for implants with a positive difference of 23 and a percentage difference of 2.65%, Male Condoms with a positive difference of 7495 and Female Condoms with a positive difference of 180.

However with a negative difference of -4051 for injectable, it shows that the number of family planning cases for injectable and the quantity of injectable family planning health commodities administered are inconsistent.

### 5.5.3 Investigating possible factors leading to the inconsistency of family planning data

An investigation is also conducted to identify the possible reasons that may have led to the inconsistency in the data. The following reasons which may have led to the inconsistency on family planning data are considered during the analysis.

* Facilities which reported on family planning cases in the DHIS2 may not have reported on the consumption of family planning health commodities in the LMIS.
* There are more family planning cases in health facilities than the quantity of family planning health commodities available in store. Health Facilities only report on the consumption of family planning health commodities supplied to them and this may led to situations where the number of family planning cases reported is higher than the consumption of family planning health commodities reported.
* A human error where health staff or data entry staff mistakenly inputted wrong figures while reporting may also have led to the inconsistency in the data.

For each month within the four months period (i.e., January 2014, February 2014, March 2014 and April 2014), the analysis tries to determine where the error may have occurred. This is done for each of the family planning services that are inconsistent.

To investigate how the first reason may have affected the inconsistency in the data, the analysis extracted health facilities which reported on family planning cases in the DHIS2 but failed to report on family planning health commodities consumed in the LMIS and then adds family planning cases whose drug consumption are not reported to calculate the sum. Finally, the percentage of cases whose consumption are not reported on the LMIS is then calculated. This analysis is done for each family planning service that is inconsistent and only considers months with a negative difference (i.e. cases > consumption).

Below is the formula to calculate the difference and percentage difference.

***Let:***

***Number of family planning cases= cases***

***Quantity of family planning health commodities administered by health facilities= consumption***

***Sum of cases of health facilities whose consumption are not reported =not reported***

***Difference=cases-consumption***

***Percentage difference=ABS (not reported/ (cases – consumption))\* 100***

To investigate whether the second reason may have resulted to the inconsistency in the data, the analysis compares the quantity of family planning health commodities supplied to health facilities to the quantity consumed by health facilities. This is done for months where there is a negative difference.

For Pujehun district, Oral Contraceptive Pills, Implants, Injectable, Male Condoms, and Female Condoms family planning services are considered. They are all considered because all of them do not satisfy the first condition to be considered consistent.

For Tonkolili district, only Injectable is considered because it is the only family planning service that does not satisfy the condition to be considered consistent.

#### 5.5.3.1 Investigating factors leading to the inconsistency of family planning data for Pujehun district.

##### 5.5.3.1.1 Oral Contraceptive Pills:

Further analysis is performed on oral contraceptive service to determine if the factors above are reasons for the inconsistency in the data.

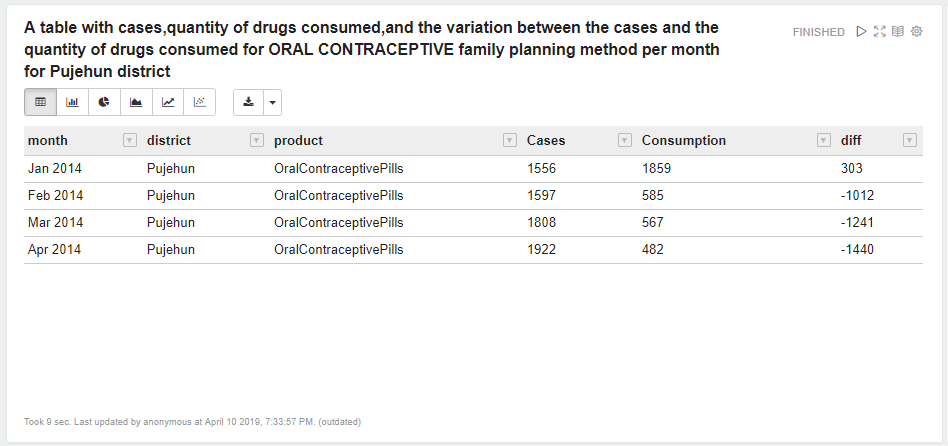


Table 5. 4 cases, consumption, and difference for Oral Contraceptive pills in Pujehun district

The table above shows that there is a negative difference between the Cases and Consumption in February 2014, March 2014 and April 2014. Therefore the investigation concludes that the error occurred during these months and therefore only considers them for further investigation.

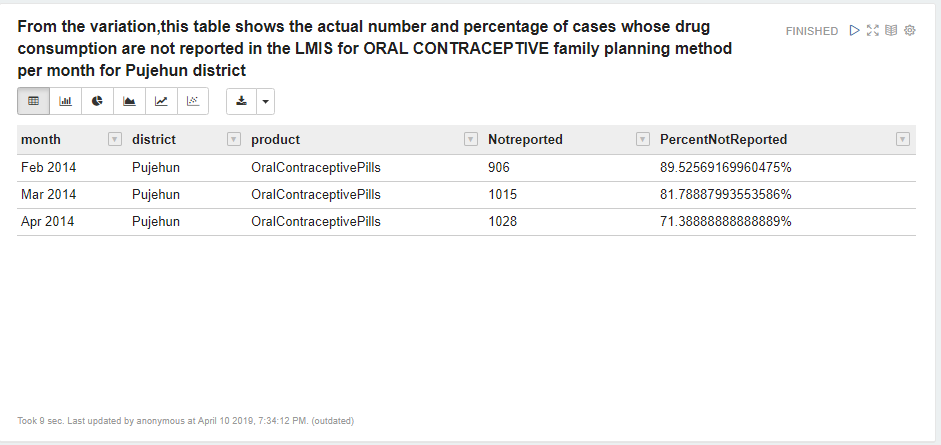


Table 5. 5 number of oral contraceptive pills cases whose consumption are not reported in the LMIS and its percentage for Pujehun district.

For January 2014, out of a difference of 1012 between the cases and the consumption, the quantity of oral contraceptive pills administered are not reported for 906 cases in the LMIS.This gives 89.5% of the difference. The same applies for February with 1015 not reported and a percentage of 81.7% and for April 2014, with 1028 not reported and a percentage of 71.39%.

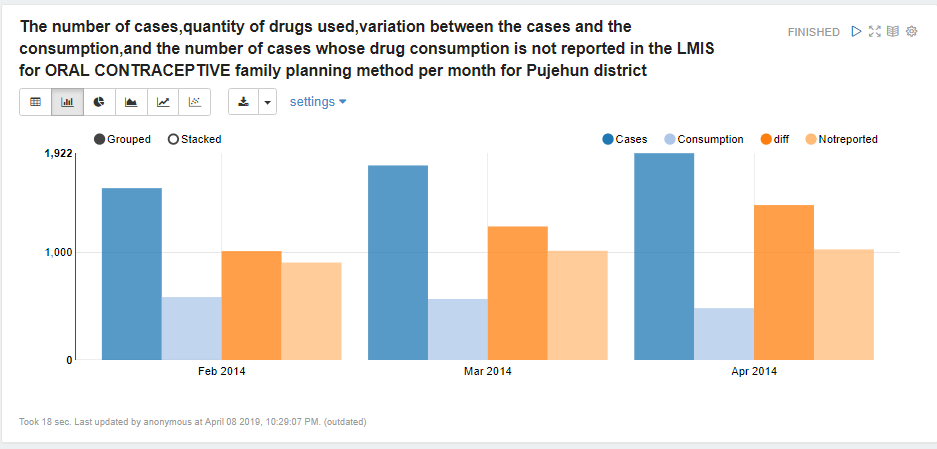


Figure 5. 4 a comparison between the oral contraceptive cases, consumption, difference and cases whose consumption are not reported for Pujehun district per month.

To investigate whether the second factor is a possible reason leading to the inconsistency in the data, the bar chart below compares the number of oral contraceptive pills health facilities had in store to what is dispensed(i.e., consumed).

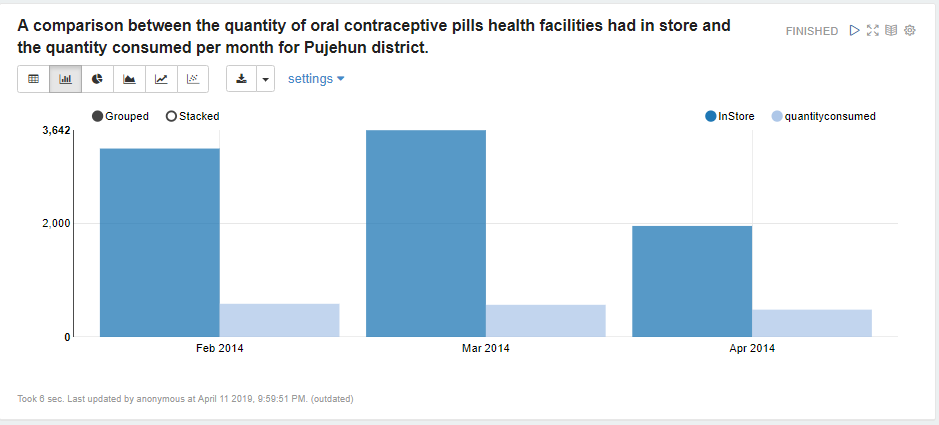


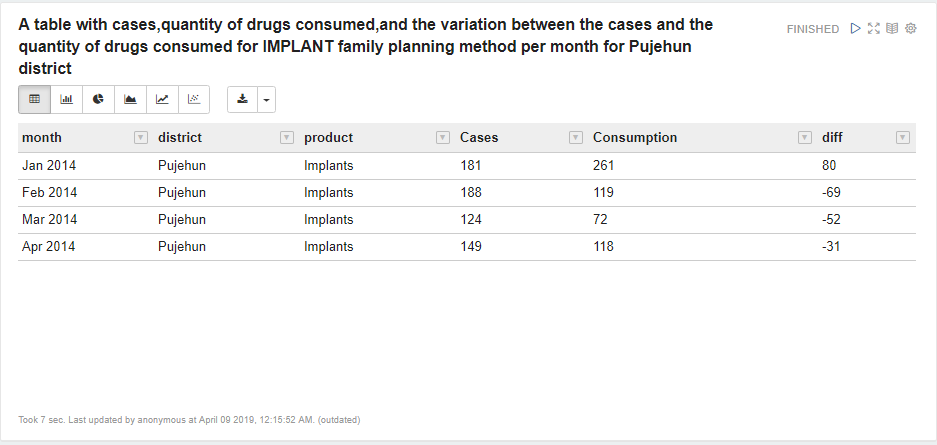
Figure 5. 5 comparison between the quantity of oral contraceptive pills in store and the quantity dispensed.

From the bar chart above, the quantity of oral contraceptive pills in store far exceeds what is dispensed(i.e., consumed). This implies that facilities always have in store more than what they consume and therefore this factor is not considered as a reason for the inconsistency in the data.

With a percentage difference of 89.5% in January, 81.79% in February and 71.39% in March we can conclude that the first reason is major cause of the inconsistency in Oral Contraceptive family planning service data.

The others may be due to the third reason or other external factors that cannot be determined from this analysis.

##### 5.5.3.1.2 Implant

 Table 5. 6 cases, consumption, and difference for Implants family planning service in Pujehun district

The table above also shows that the inconsistency between the cases and consumption occurred in February 2014, March 2014 and April 2014.

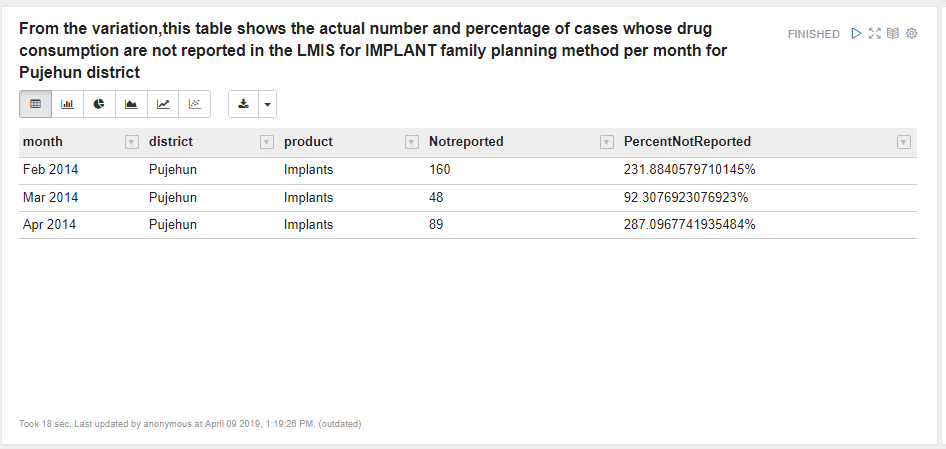


Table 5. 7 number of implant cases whose consumption are not reported in the LMIS and its percentage for Pujehun district.

From the table above, we can see that although the difference between the cases and consumption for February 2014 is 69, however, it returns a value greater than the difference(i.e., 160) for cases whose consumption are not reported(i.e., Notreported). This happens because facility names from the three data sources are not spelt the same way. Although they may mean the same facility but the names are spelt differently. Therefore consumption for cases of these facilities may have been reported, but the query sees it as if they are not reported because the names are spelt differently.

Below is the SQL query which determines cases of facilities whose consumption are not reported.

***CREATE TABLE fp\_notin\_jan AS***

***SELECT fp\_dhis\_inc\_Jan.district,fp\_dhis\_inc\_Jan.facility,fp\_dhis\_inc\_Jan.date,fp\_dhis\_inc\_Jan.product,fp\_dhis\_inc\_Jan.value***

***FROM fp\_dhis\_inc\_Jan***

***LEFT JOIN fp\_lmis\_inc\_Jan ON (SUBSTRING(fp\_dhis\_inc\_Jan.facility,1,5)) = (SUBSTRING(fp\_lmis\_inc\_Jan.facility,1,5))***

***WHERE SUBSTRING(fp\_lmis\_inc\_Jan.facility,1,5) IS NULL***

***GROUP BY fp\_dhis\_inc\_jan.facility,fp\_dhis\_inc\_jan.district,fp\_dhis\_inc\_jan.product,fp\_dhis\_inc\_jan.value,fp\_dhis\_inc\_Jan.date***

***ORDER BY fp\_dhis\_inc\_jan.district ASC***

The Hive query language above compares the first five characters of the names of facilities in both the DHIS2 and LMIS data sources and then return facility names whose cases are reported(i.e. from the DHIS2 ) but the consumption of health commodities for those cases are not reported(i.e. from the LMIS).

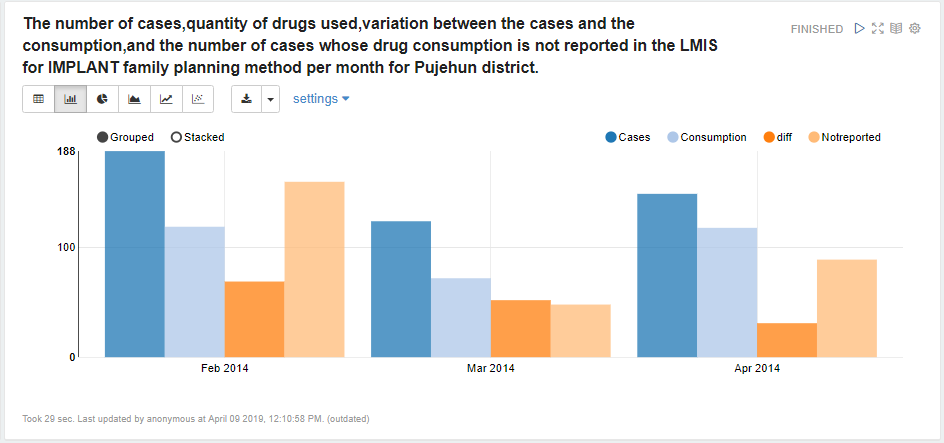
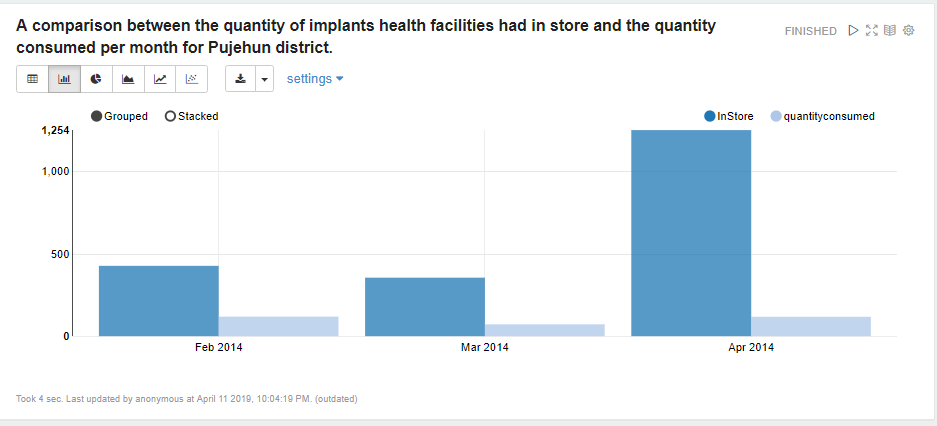


Figure 5. 6 a comparison between implant cases, consumption, difference and number of implant cases whose consumption are not reported for Pujehun district per month.

To also investigate whether the second reason is a factor which leads to the inconsistency on implant family planning data, the figure below compares the quantity of implants health facilities had in store to what they consumed within that period.

 Figure 5. 7 comparison between the quantity of Jadelle implants in store and the quantity dispensed.

What health facilities had in store far exceeds what is consumed. Therefore the second reason is also not a factor which resulted to the inconsistency of Implants family planning data.

##### 5.5.3.1.3 Injectables

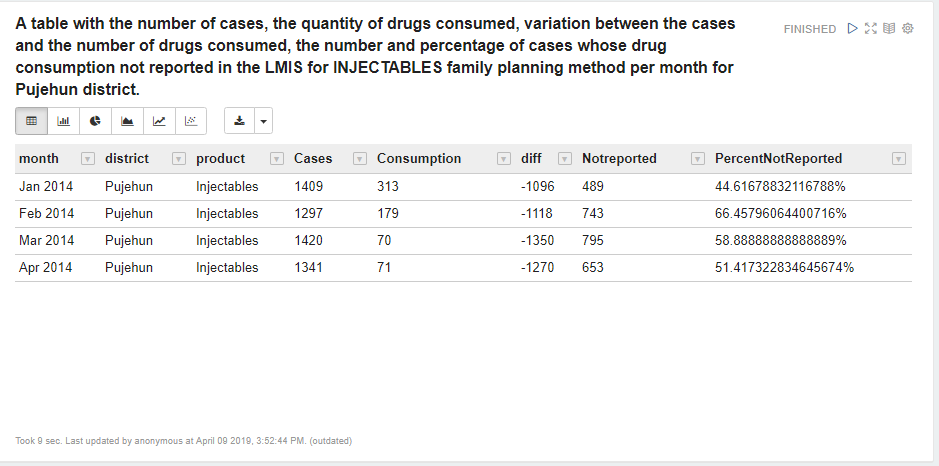


Table 5. 8 cases, consumption, difference, number of cases for injectable whose consumption are not reported in the LMIS and its percentage in Pujehun district.

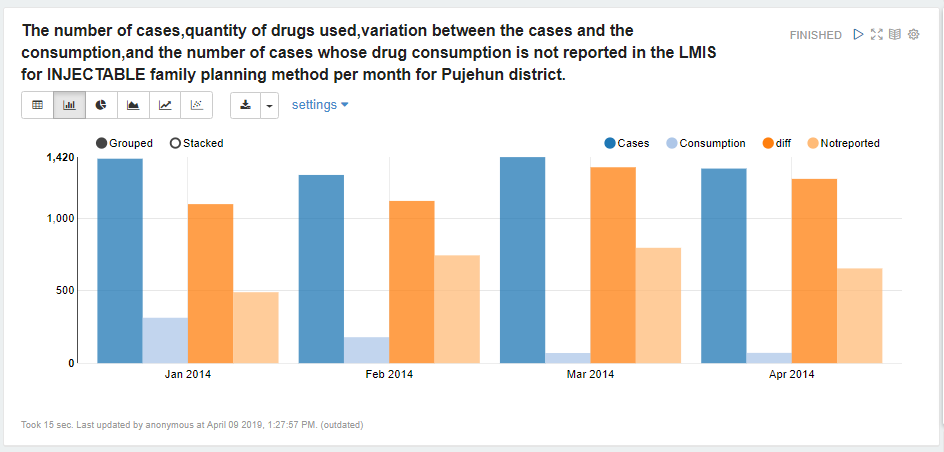


Figure 5. 8 comparison between injectable cases, consumption, difference, and number of injectable cases whose consumption are not reported for Pujehun district.

From the figures above, we can deduce that out of a difference of 1096 between the number of cases for injectable family planning services and the consumption of injectable family planning health commodities, there are 489 cases whose consumption are not reported in the LMIS. This gives a percentage of 44.6% for January 2014.The same applies for February 2014 with 743 cases whose drug consumption are not reported and this also gives 66.5%.

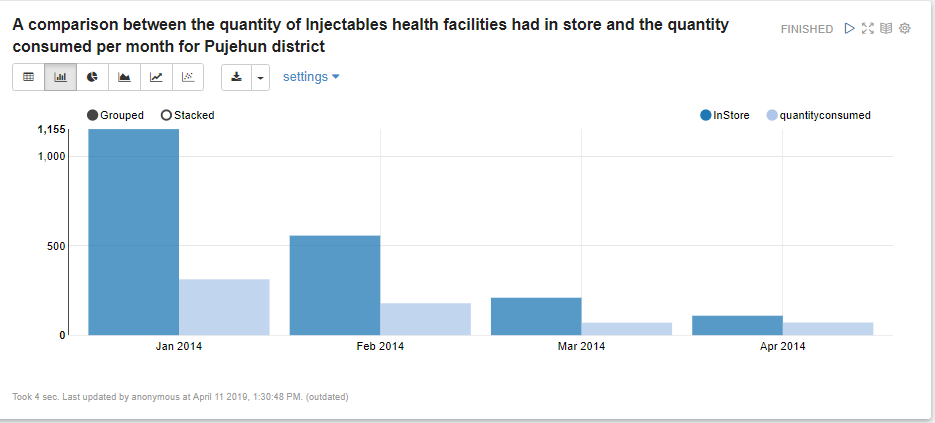


Figure 5. 9 comparison between the quantity of Injectable (i.e., Depo Provera) in store and the quantity dispensed.

Figure 5.8 shows that for each month, the quantity of injectable (i.e., Depo Provera) health facilities had in store exceeds the quantity consumed by health facilities.

From these analyses, we can say that the first reason is also one of the major causes of the inconsistency in injectable family planning data. Other reasons such as human error and external factors which cannot be determined from this analysis may also have been another cause.

##### 5.5.3.1.4 Male Condoms

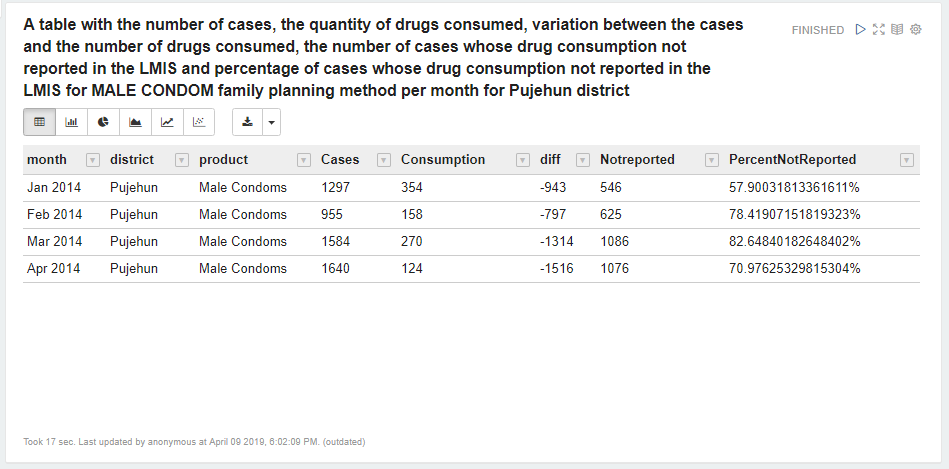


Table 5. 9 cases, consumption, difference, and cases whose consumption are not reported in the LMIS and its percentage for male Condoms in Pujehun district.

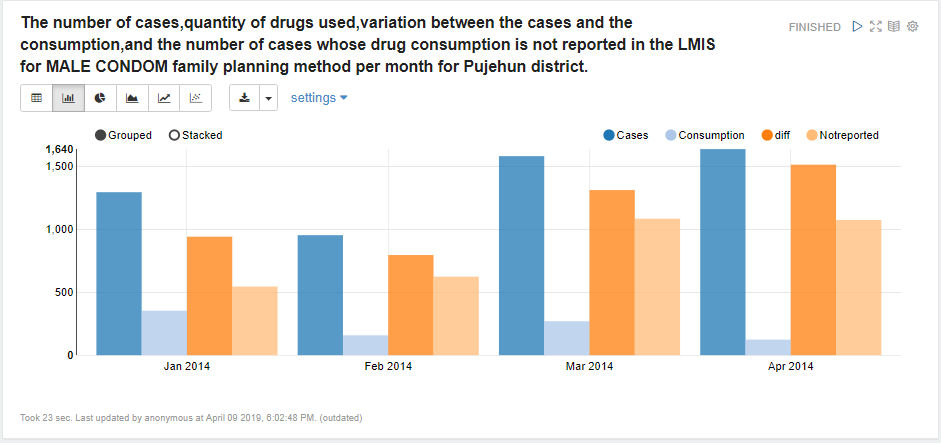


Figure 5. 10 comparison between cases, consumption, difference, and number of cases whose consumption are not reported for male condoms in Pujehun district.

From the figures above, there is a percentage of 57.9%, 78.4%, 82.6% and 70.98% for cases whose drug consumption are not reported for the periods of January, February, March, and April respectively.

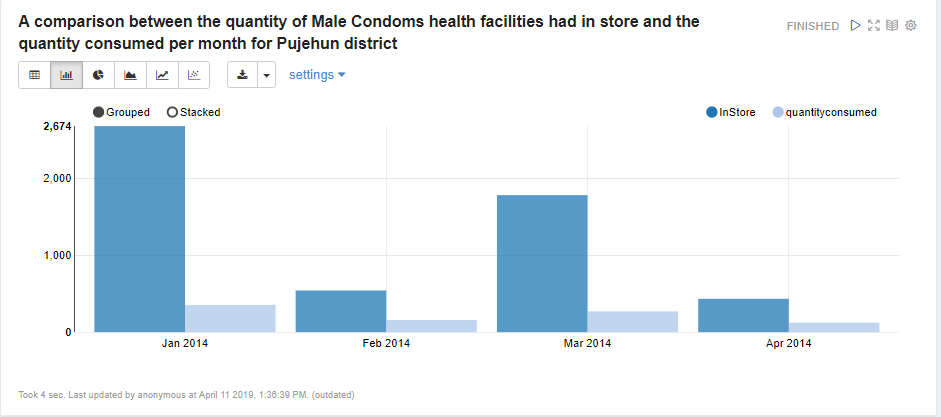


Figure 5. 11 comparison between the number of male condoms in store and the quantity dispensed.

Figure 5.10 above shows that the quantity of male condoms health facilities had in store far exceeds what they consume for January, February, March, and April respectively.

From this analysis, we can say that the first factor is a major cause of the inconsistency on Male Condoms family planning data.

##### 5.5.3.1.5 Female Condoms

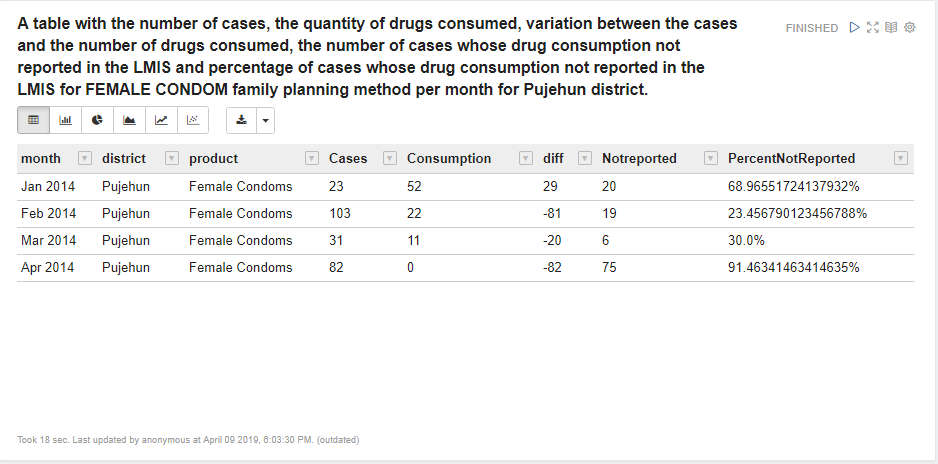


Table 5. 10 cases, consumption, difference, and cases whose consumption are not reported in the LMIS and its percentage for female Condoms in Pujehun district.

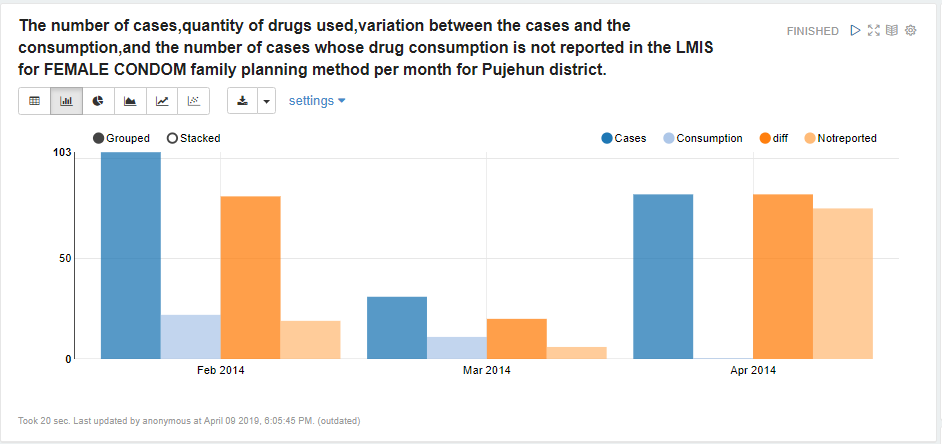


Figure 5. 12 comparison between cases, consumption, difference, and number of cases whose consumption are not reported for female condoms in Pujehun district.

The figures above show that the error (i.e., where cases > consumption) occurred in February, March, and April. For April 2014, the figures shows that only about 8.5% of the consumption was reported for cases.

Although there is a positive difference for January, it still shows that there are some facilities which reported about Female Condoms cases on the DHIS2 but failed to report on the consumption of Female Condoms for cases reported. This is a reoccurrence of what happened for implants family planning service.



Figure 5. 13 comparison between the number of male condoms in store and the quantity dispensed.

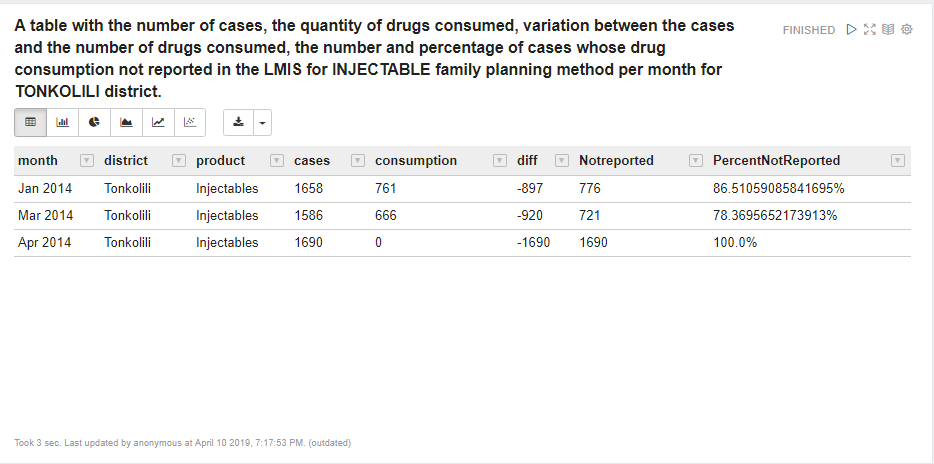
Figure 5.13 above shows that the quantity of female condoms facilities had in store exceed the quantity consumed. Therefore the second factor is not a cause of the inconsistency on female condoms family planning data.

#### 5.5.3.2 Investigating factors leading to the inconsistency of family planning data for Tonkolili district.

##### 5.5.3.2.1. Injectable

For Tonkolili district, there is a negative difference for only injectable family planning service, and therefore it is considered to be inconsistent.

The figures below show an analysis of the possible reasons that may have resulted in the inconsistency of family planning data.

Table 5. 11 cases, consumption, difference, cases whose consumption are not reported in the LMIS and its percentage for Injectable in Pujehun district.

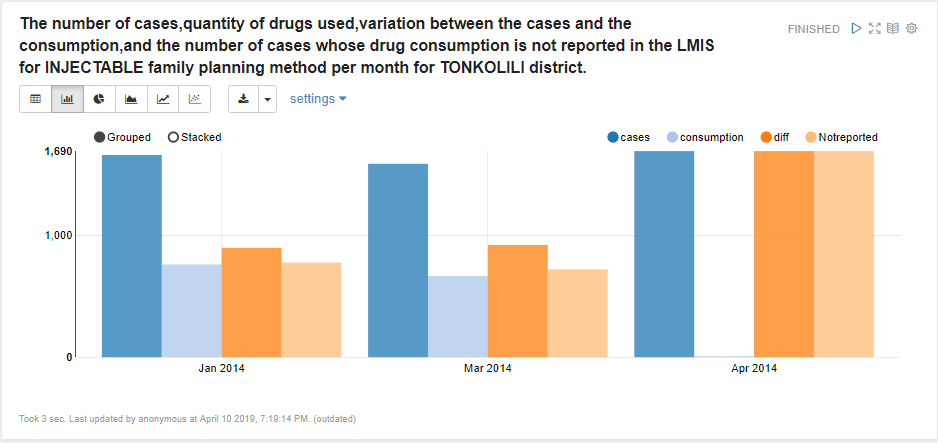


Figure 5. 14 comparison between cases, consumption, difference, and number of cases whose consumption are not reported for Injectable in Pujehun district.

The figures above show that the negative variation between the cases and consumption for Injectables occurs in January, February, and March. No report was made on the consumption for April although a report on the cases was made. Therefore, this gives a percentage difference of 100% for April.



Figure 5. 15 comparison between the quantity of Injectable in store and the quantity dispensed.

From figure 5.12 above, it shows that no report is made for Injectable for April. For January and February, the quantity in store exceeds the quantity consumed.

## 5.6 Challenges

Although this work is successful as the main aim is to build an interactive platform which integrates data from multiple sources, stores it in a single repository and ensures analysis and visualization of the data for the ministry of health and sanitation in Sierra Leone, however, few challenges were also experienced in carrying out this work.

The first challenge has to do with resource constraints. Big Data technologies require hardware with huge storage and high processing power. This makes it very challenging to build a single big data platform which integrates data from multiple sources, store the data and ensures analysis using different Big Data tools. This, however, resulted in some limitations on the platform when it comes to ingesting very large data sets, perform complex data analysis and configuration.

Secondly, building such a platform requires the use of multiple tools and understanding all of the tools is time consuming. Four different tools are used for this work and therefore requires extensive research on each of these tools to fully understand their functionalities and how suitable are they for such project.

Finally, during the analysis, it is observed for some facilities, that the names are spelt differently on each of the data sources. This interferes with the accuracy of the analysis when it comes to comparing attributes that are common among the three data sources.

An example here is a case where the health facility name is used as a common attribute from the two data sources to compare the family planning cases and the consumption of family planning health commodities. Such effect is observed in the case of implant for Pujehun district where a percentage of over 100% is returned. It is also observed for Female condoms in Pujehun district where it returned facilities that did not report on the consumption of female condoms although the difference between the cases and the consumption is positive. This occurs because the query compares the first five characters of health facilities in both the DHIS2 and the LMIS and this becomes a challenge in cases where the names of some health facilities are spelt differently in the two data sources.

Below is the query in Spark SQL.

***CREATE TABLE fp\_notin\_jan AS***

***SELECT fp\_dhis\_inc\_Jan.district,fp\_dhis\_inc\_Jan.facility,fp\_dhis\_inc\_Jan.date,fp\_dhis\_inc\_Jan.product,fp\_dhis\_inc\_Jan.value***

***FROM fp\_dhis\_inc\_Jan***

***LEFT JOIN fp\_lmis\_inc\_Jan ON (SUBSTRING(fp\_dhis\_inc\_Jan.facility,1,5)) = (SUBSTRING(fp\_lmis\_inc\_Jan.facility,1,5))***

***WHERE SUBSTRING(fp\_lmis\_inc\_Jan.facility,1,5) IS NULL***

***GROUP BY fp\_dhis\_inc\_jan.facility,fp\_dhis\_inc\_jan.district,fp\_dhis\_inc\_jan.product,fp\_dhis\_inc\_jan.value,fp\_dhis\_inc\_Jan.date***

***ORDER BY fp\_dhis\_inc\_jan.district ASC***

The Hive Query language above creates a table, select records from different tables in both the DHIS2 and LMIS where facilities reported in the DHIS2 are not reported in the LMIS and then insert the data to the table created.

## 5.7 Conclusion

The analysis above has demonstrated that is important to have a platform which automatically extracts data from external data sources and ingest it to a repository, stores the data and ensures data analysis and visualization. With this, data integration, data storage, data selection and data analysis are all done in a single integrated platform without facing the difficulties of manually moving data from one location to the other. This single integrated platform enables data from different data sources to be thoroughly explored in a single location to extract knowledge from the data. The integrated platform ensures that data from different data sources is readily available almost on a real time basis. This enables data analyst or data scientists to collaborate and compare data from different data sources to make meaningful judgments which will lead to effective and efficient data analysis. It also ensures prompt decision making and enables the ministry to have a clear understanding of what is happening.

Secondly, the analysis also discovered that there is a high level of inconsistency of family planning data between the number of people who visited health facilities for family planning services (i.e., Family planning cases) and the consumption of family planning health commodities for the cases reported. The analysis went further to investigate the major causes of the inconsistency in the data. The analysis then concludes that a major causes is a failure on the part of the health facilities to report on the consumption of family planning health commodities in the LMIS for cases that have already been reported in the DHIS2. Other factors such as what facilities had in store and the consumption of health commodities is not a cause of the inconsistency in the data from the analysis. Human error such as typographical errors and other external factors may also be a reason for the inconsistency in the data.

Therefore this work clearly justifies the importantance of having an interactive platform which integrates data from multiple sources because this will enable the ministry of health and sanitation to understand what is happening,identitfy the root cause of the problems and then take the approriate actions.

# Acknowledgement

Firstly, I acknowledge the guidance and protection of the almighty God throughout this work.

Secondly, my sincere gratitude goes to my supervisor Shu Liu for his guidance and advice. Thank you for your support and inspiration sir.

Finally, a special thanks go to my darling wife Bintu Fofanah for your words of encouragement throughout these difficult times. You have always been there for me, and I am grateful for your words of encouragement. I LOVE YOU

# REFERENCES

[1]Jasmine Zakir, Tom Seymour and Kristi Berg.”Big Data Analytics”, Issues in Information Systems, Volume 16, 2015, 81-90

[2] Zaiying Liu, Ping Yang and Lixiao Zhang.A Sketch of Big Data Technologies,Seventh International Conference on Internet Computing for Engineering and Science,2013

[3] Transparency Market Research.Global Big Data Market Driven by Increasing Demand from Government Healthcare BFSI Sectors.2016. [Online] Available:https://www.transparencymarketresearch.com/pressrelease/global-big-data-market.htm.

[4] Vikash Yadav, Monika Verma and Vandana Dixit Kaushik.Big Data Analytics for Health Systems.International Conference on Green Computing and Internet of Things (ICGCloT),2015

[5] Sonja Zillner, Heiner Oberkampf, Claudia Bretschneider, Amrapali Zaveri, Werner Faix and Sabrina Neururer. Towards a Technology Roadmap for Big Data Applications in the Healthcare Domain.IEEE IRI 2014, San Francisco, California, USA, August 13-15, 2014

[6] Umar Ahsan and Abdul Bais, A Review on Big Data Analysis and Internet of Things, IEEE 13th International Conference on Mobile Ad Hoc and Sensor Systems. 2016

[7] Han Hu, Yonggang, Wen, TAT-Seng Chua and Xuelong Li .Toward Scalable Systems for Big Data Analytics. IEEE.2014

[8]Brief history of Big Data. Journal by Anastasia.29 August 2016. [Online] Available: <https://www.cleverism.com/brief-history-big-data/>.

[9] Megan Sheeran and Robert Steele “A Framework for Big Data Technology in Health and Healthcare”, IEEE, 2017

[10] D.Singh and C.K Reddy.A survey on platforms for big data analytics.2014

[11] Apache Zookeeper. [Online] Available: <https://zookeeper.apache.org>. [Retrieved October 2018]

[12] B.Saraldevi, N.Pazhaniraja,P.Victor Paul,M.S Saleem Basha and P.Dhavachelvan.Big Data and Hadoop-A Study in Security Perspective. 2nd International Symposium on Big Data and Cloud Computing.2015

[13] Shyram R,Bharathi Gannesh HB,Sachin Kumar S,Prabaharan Poornachan dran and Soman KP.Apache Spark a Big Data Analytics Platform for Smart Grid.Smart Grid Technologies,August 6-8,2015

[14] Andreas P. Plageras, Christos Stergiou, GeorgeKokkonis, Kostas E. Psannis, Yutaka Ishibashi, Byung-Gyu Kim and B. Brij Gupta.Efficient Large-scale Medical Data (eHealth Big Data) Analytics in Internet of Things.IEEE 19th Conference on Business Informatics.2017

[15] Jing Lu and Malcolm Keech.Emerging Technologies for Health Data Analytics Research: A Conceptual Architecture.26th International Workshop on Database and Expert Systems Applications.2015

[16] P. Gonzalez-Alonso, R. Vilar and F. Lupiañez-Villanueva.Meeting Technology and Methodology into Health Big Data Analytics Scenarios.IEEE 30th International Symposium on Computer-Based Medical Systems.2017

[17] National Health Sector Strategic Plan. Ministry Of Health and Sanitation. September, 2017. [Online] Available: <http://www.nationalplanningcycles.org/sites/default/files/planning_cycle_repository/sierra_leone/sierra_leone_nhssp_2017-21_final_sept2017.pdf>. [Retrieved October 2018]

[18] Sierra Leone National Health Information System Strategic Plan (2017 -2021).Ministry Of Health and Sanitation, [Online] Available: <https://mohs2017.files.wordpress.com/2017/06/his-strategic-plan-2017-2021.docx>. (Retrieved October 2018)

[19] Sierra Leone Human Resources for Health Strategic Plan (2017-2021).Ministry Of Health and Sanitation. [Online]Available: <http://www.afro.who.int/publications/sierra-leone-human-resources-health-strategic-plan-2017-2021>. [Retrieved October 2018]

[20] Horton Works Data Platform.Hortonworks. [Online] Available: <https://hortonworks.com/products/data-platforms/hdp/>. [Retrieved December 2018]

[21] District Health Information Software. [Online] Available: <https://www.dhis2.org/>. [Retrieved December 2018]

[22] Integrated Human Resource Information System (iHris). [Online] Available: <https://www.ihris.org/ihris-suite/health-workforce-software/ihris-manage/>. [Retrieved December 2018]

[23] Sangwhan Cha, Ashraf Abusharekh and Syed SR Abidi.Towards a Big Health Data Analytics platform. IEEE First International Conference on Big Data Computing Service and Applications.2015

[24] Anirudh Kadadi, Rajeev Agrawal, Christopher Nyamful and Rahman Atiq.Challenges of Data Integration and Interoperability in Big Data**.**2014 IEEE International Conference on Big Data.2014

[25] Defining Architecture Components of the Big Data Ecosystem”. System and Network Engineering Group,University of Amsterdam.Hong Kong Polytechnic University,2014

[26] BiswaRanjan Samal1, Mrutyunjaya Panda.Real Time Product Feedback Review and Analysis Using Apache Technologies and NOSQL Database, P.G Department of Computer Science and Application, Utkal University,2017

[27]Apache Nifi Overview. [Online] Available:https://nifi.apache.org/docs/nifi-docs/html/overview.html.[Retrieved April 2019]

[28]Apache Hadoop.Hortonworks. [Online] Available:https://hortonworks.com/apache/hadoop/. [Retrieved March 2019]

[29]Tracking JobTracker and TaskTracker in Hadoop1, [Online] Available:https://www.dummies.com/programming/big-data/hadoop/tracking-jobtracker-and-tasktracker-in-hadoop-1/.[Retrieved March 2019]

[30] Yuri Demchenko, Cees de Laat and Peter Membrey.Defining Architecture Components of the Big Data Ecosystem, IEEE, 2014

[ 31]Spark Interpreter for Apache Zeppelin. Zeppelin. [Online] Available:https://zeppelin.apache.org/docs/latest/interpreter/spark.html. [Retrieved March 2019]

[32]HDFS User Guide.[Online] Available:<https://hadoop.apache.org/docs/stable/hadoop-project-dist/hadoop-hdfs/HdfsUserGuide.html>. [Retrieved March 2019]

[33] Apache Hadoop | Hadoop Architecture - Altencalsof, [Online] Available:https://www.altencalsoftlabs.com/apache-hadoop/ [Retrieved April 06, 2019].

[34] Hdfs Architecture Guide - Apache Hadoop, [Online] Available:https://hadoop.apache.org/docs/r1.2.1/hdfs\_design.pdf [Retrieved April 06, 2019].

[35] Pak Dba: Working with Apache Spark Sql, [Online] Available:https://pakdba.blogspot.com/2017/12/working-with-apache-spark-sql.html [Retrieved April 06, 2019].

[36]Family Planning/Contraception. World Health Organization. [Online] Available: <https://www.who.int/news-room/fact-sheets/detail/family-planning-contraception>. [Retrieved April 2019]

[37]What are the different methods of birth control-Family Planning Council. [Online] Available:<https://www.fpcouncil.com/content/faq/what-are-different-methods-birth-control>[Retrieved April 2019]

[38]Family Planning Methods. [Online] Available:<https://www.fphandbook.org/sites/default/files/fffpchapter7.pdf>.[Retrieved April 2019]

[39] Ardi Imawan and Joonho Kwon.A Timeline Visualization System for Road Traffic Big Data. 2015 IEEE International Conference on Big Data (Big Data). 2015

[40] Hortonworks Data Platform.Hortonworks [Online] Available:https://hortonworks.com/products/data-platforms/hdp/ [Retrieved April 2019]

[41]Apache Ambari.Hortonworks. [Online] Available:https://hortonworks.com/apache/ambari/#section\_1.[Retrieved April 2019]

[42] Cloudera Documentation.Cloudera. [Online] Available:https://docs.hortonworks.com.[Retrieved April 2019]

[43] Kyounghyun Park, Minh Chau Nguyen and Heesun Won. Web-based Collaborative Big Data Analytics on Big Data as a Service Platform. Big Data SW Research Department, Electronics and Telecommunication Research Institute, South Korea. July 1-3, 2015

[44] Ruimin Hu.Key Technology for Big Visual Data Analysis in Security Space and its Applications.2016 International Conference on Advanced Cloud and Big Data.2016

[45] Meryeme EL HOUARI, Maryem RHANOUI and Bouchra EL ASRI.Hybrid Big Data Warehouse for On-demand decision needs.3rd International Conference on Electrical and Information Technologies ICEIT’2017. 2017

[46] Alka Londhe and PVRD Prasada Rao.Platforms for Big Data Analytics: *Trend towards Hybrid Era.*International Conference on Energy, Communication, Data Analytics and Soft Computing (ICECDS-2017). 2017

[47] Kezia Rani.B and Vinaya Babu.A.Scheduling of Big Data Application Workflows in Cloud and Inter-Cloud Environments. 2015 IEEE International Conference on Big Data (Big Data). 2015

# Appendix Abbreviations List

Acronyms Full Meaning

DHMT District Health Management Team

LMIS Logistics Management Information System

DHIS2 District Health Information System 2

IHRIS Integrated Human Resource Information System

ORC Optimized Row Columnar

RC Record Columnar

UI User Interface

1. Nguyen, Van-Quyet and Nguyen, Sinh Ngoc and Kim, Kyungbaek, Design of a Platform for Collecting and Analyzing Agricultural Big Data [↑](#endnote-ref-1)
2. Barghash, Shadi, Big Data Platform for Agriculture Bachelor Thesis [↑](#endnote-ref-2)
3. Cha, Sangwhan and Abusharekh, Ashraf and Abidi, Syed S.R., Towards a 'big' health data analytics platform [↑](#endnote-ref-3)
4. Reddy, D. Siva Sankara and Kumar, R. Udaya, Big Data Analytics for Healthcare Organization: A Study of BDA Process, Benefits and Challenges of BDA. [↑](#endnote-ref-4)
5. Drakos, Andreas and Protonotarios, Vassilis and Manouselis, Nikos, agINFRA: A research data hub for agriculture, food and the environment [↑](#endnote-ref-5)
6. Stubbs, Megan, Big Data in U.S. Agriculture [↑](#endnote-ref-6)
7. https://www.gartner.com/newsroom/id/492112 [↑](#endnote-ref-7)
8. Ungerer, Gerhard, Cleaning Up the Data Lake with an Operational Data Hub. [↑](#endnote-ref-8)
9. Paper, Research-in-progress and Warehousing, Data and Intelligence, Business, From Data Warehouses To Transformation Hubs [↑](#endnote-ref-9)
10. Big Data, NATHAN MARZ, JAMES WARREN [↑](#endnote-ref-10)
11. Big Data Applications and Use Cases, Patrick C. K. Hung [↑](#endnote-ref-11)
12. BIG DATA FOR AGRICULTURE, Gerard Sylvester [↑](#endnote-ref-12)
13. Big data demystified, David Stephenson [↑](#endnote-ref-13)
14. Hadoop in Practice, Alex Holmes [↑](#endnote-ref-14)
15. Hadoop in Action, Chuck Lam [↑](#endnote-ref-15)
16. Hadoop: What You Need to Know, Donald Miner [↑](#endnote-ref-16)
17. Programming Hive, Edward Capriolo, Dean Wampler, and Jason Rutherglen [↑](#endnote-ref-17)
18. Apache Hive Essentials, Dayong Du [↑](#endnote-ref-18)
19. Apache Nifi Overview. [Online] Available:https://nifi.apache.org/docs/nifi-docs/html/overview.html.[Retrieved April 2020] [↑](#endnote-ref-19)
20. **Learning Spark SQL,** Aurobindo Sarkar [↑](#endnote-ref-20)
21. **Spark: The Definitive Guide,** **Bill Chambers, and Matei Zaharia** [↑](#endnote-ref-21)
22. **https://spark.apache.org/sql/** [↑](#endnote-ref-22)
23. https://zeppelin.apache.org/docs/0.6.0/ [↑](#endnote-ref-23)
24. https://www.cloudera.com/products/open-source/apache-hadoop/key-cdh-components.html [↑](#endnote-ref-24)
25. https://www.cloudera.com/downloads/cdh.html [↑](#endnote-ref-25)
26. https://docs.cloudera.com/documentation/enterprise/5-9-x/topics/cdh\_intro.html [↑](#endnote-ref-26)
27. BIG DATA FOR AGRICULTURE, Gerard Sylvester [↑](#endnote-ref-27)
28. <https://bigdata.cgiar.org/about-the-platform/> [↑](#endnote-ref-28)
29. Leveraging CGIAR data, CGIAR [↑](#endnote-ref-29)
30. http://www.fao.org/waicent/faostat/agricult/pr\_ele-e.htm [↑](#endnote-ref-30)
31. https://www.ers.usda.gov/data-products/food-availability-per-capita-data-system/glossary/ [↑](#endnote-ref-31)
32. https://dictionary.cambridge.org/us/dictionary/english/domestic-consumption [↑](#endnote-ref-32)
33. Apache Nifi Overview. [Online] Available:https://nifi.apache.org/docs/nifi-docs/html/overview.html.[Retrieved April 2020] [↑](#endnote-ref-33)
34. Tracking JobTracker and TaskTracker in Hadoop1, [Online] Available:https://www.dummies.com/programming/big-data/hadoop/tracking-jobtracker-and-tasktracker-in-hadoop-1/.[Retrieved March 2020] [↑](#endnote-ref-34)