High Level Design (HLD) for

Shaw – Hortonworks Architecture



Revision History

|  |  |  |  |
| --- | --- | --- | --- |
| **Date** | **Version** | **Author(s)** | **Remarks** |
| 7/29/2016 | 1.0 | Scott Crystal | Initial Version |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |

Document References

|  |  |  |  |
| --- | --- | --- | --- |
|  |  |  |  |
| **Date** | **Version** | **Author(s)** | **Document/File Title** |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |

Table of Contents

[1.0 Project Introduction 1](#_Toc457687435)

[1.1 Overview 1](#_Toc457687436)

[2.0 Apache Hadoop 2](#_Toc457687437)

[2.1 Hortonworks 2](#_Toc457687438)

[2.1.1 Administration 2](#_Toc457687439)

[2.1.2 Batch Processing 2](#_Toc457687440)

[2.1.3 Streaming 3](#_Toc457687441)

[2.1.4 Scheduling 3](#_Toc457687442)

[2.1.5 Database 4](#_Toc457687443)

[3.0 Master Data Management (MDM) 5](#_Toc457687444)

[3.1 Naming Standards 5](#_Toc457687445)

[4.0 Security 6](#_Toc457687446)

[5.0 Reporting 7](#_Toc457687447)

[5.1 Tableau 7](#_Toc457687448)

[6.0 Appendix A – Acronyms List 8](#_Toc457687449)

[7.0 Appendix B – 9](#_Toc457687450)

List of Tables

[Table 4‑1: List of Data Sources 5](#_Toc457688348)

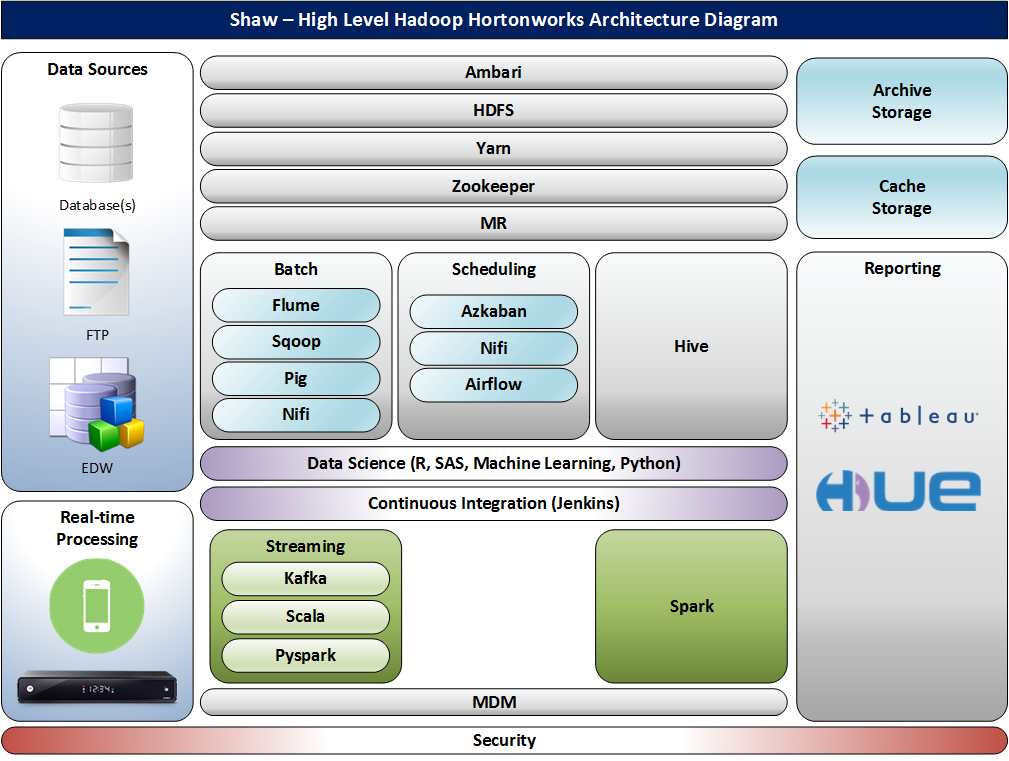
List of Figures

[Figure 1–1: Shaw – High Level Hadoop Hortonworks Architecture Diagram 1](#_Toc457681384)

# Project Introduction

## Overview

Figure 1–1: Shaw – High Level Hadoop Hortonworks Architecture Diagram



# Apache Hadoop

The Apache Hadoop project develops open-source software for reliable, scalable, distributed computing.

The Apache Hadoop software library is a framework that allows for the distributed processing of large data sets across clusters of computers using simple programming models. It is designed to scale up from single servers to thousands of machines, each offering local computation and storage. Rather than rely on hardware to deliver high-availability, the library itself is designed to detect and handle failures at the application layer, so delivering a highly-available service on top of a cluster of computers, each of which may be prone to failures. (<http://hadoop.apache.org/>)

## Hortonworks

Hortonworks, Inc. is a leading innovator in the data industry, creating, distributing and supporting enterprise-ready open data platforms and modern data applications (<http://hortonworks.com/>)

### Administration

#### HDFS

Hadoop Distributed File System (HDFS) is a distributed file system that provides high-throughput access to application data.

#### Yarn

Yarn is a framework for job scheduling and cluster resource management.

#### MapReduce

MapReduce is a YARN-based system for parallel processing of large data sets.

#### Ambari

Ambari is a web-based tool for provisioning, managing, and monitoring Apache Hadoop clusters which includes support for Hadoop HDFS, Hadoop MapReduce, Hive, HCatalog, HBase, ZooKeeper, Oozie, Pig and Sqoop. Ambari also provides a dashboard for viewing cluster health such as heatmaps and ability to view MapReduce, Pig and Hive applications visually alongwith features to diagnose their performance characteristics in a user-friendly manner. (<http://ambari.apache.org/>)

#### Zookeeper

Zookeeper is a high-performance coordination service for distributed applications. (<http://zookeeper.apache.org/>)

### Batch Processing

#### Flume

Flume is a distributed, reliable, and available service for efficiently collecting, aggregating, and moving large amounts of log data. It has a simple and flexible architecture based on streaming data flows. It is robust and fault tolerant with tunable reliability mechanisms and many failover and recovery mechanisms. It uses a simple extensible data model that allows for online analytic application.

(<https://flume.apache.org/>)

#### Sqoop

Sqoop is a command-line interface application for transferring data between relational databases and Hadoop. (<http://hortonworks.com/apache/sqoop/>)

#### Pig

Pig is a high-level data-flow language and execution framework for parallel computation. (<http://pig.apache.org/>)

#### Nifi

Nifi is an easy to use, powerful, and reliable system to process and distribute data between systems.

(<https://nifi.apache.org/>)

### Streaming

#### Kafka

Apache Kafka is publish-subscribe messaging rethought as a distributed commit log. (<http://kafka.apache.org/>)

#### Scala

Scala is a general-purpose programming language. Scala has full support for functional programming and a very strong static type system. <https://en.wikipedia.org/wiki/Scala_(programming_language)>

(<http://www.scala-lang.org/>)

### Scheduling

#### Azkaban

Azkaban Hadoop is an open-source workflow engine for Hadoop eco systems. It is a batch job scheduler allowing developers to control job execution inside Java and especially Hadoop projects.

#### Nifi

NiFi was built to automate the flow of data between systems.

(<https://nifi.apache.org/>)

##### HORTONWORKS DATAFLOW (HDF)

Powered by Apache NiFi, Kafka and Storm, HDF collects, curates, analyzes and delivers real-time data from the IoAT to data stores both on-premises and in the cloud.

#### Airflow

Airflow is an Open Source Platform to Author and Monitor Data Pipelines.

(<http://www.slideshare.net/Hadoop_Summit/airflow-an-open-source-platform-to-author-and-monitor-data-pipelines>)

### Database

#### Hive

Hive is a data warehouse infrastructure that provides data summarization and ad hoc querying. (<http://hive.apache.org/>)

##### Hue

Hue is an open source Web interface for analyzing data with any Apache Hadoop.

<http://gethue.com/>

<https://en.wikipedia.org/wiki/Hue_(Hadoop)>

##### HBase

HBase is a scalable, distributed database that supports structured data storage for large tables. (<http://hbase.apache.org/>)

#### Spark

A fast and general compute engine for Hadoop data. Spark provides a simple and expressive programming model that supports a wide range of applications, including ETL, machine learning, stream processing, and graph computation. (<http://spark.apache.org/>)

# Cluster Configuration

## Master Node

## Gateway Node

## Data Nodes

# Data Sources

Table 4‑1: List of Data Sources

|  |  |  |
| --- | --- | --- |
| Data Source | Contact | Notes |
| Set-top Box |  |  |
| VOD Usage |  |  |
| FreeRange |  |  |
| Shaw.ca Website |  | Tagging needed |
| Adobe Analytics / Authenticated Apps data |  | Tagging needed |
| Network / TV Issues |  |  |
| Competitive Pricing and Product |  |  |
| Customer Information |  |  |
| Call Center |  |  |
| Tech Service data |  |  |
| Social Media |  |  |
| Wireless |  |  |
| Voice (VOC) |  |  |

# Data Science

## R

## SAS

## Machine Learning

## Python

# Continuous Integration

## Jenkins

# FTP

# Reporting

# Tableau

# Master Data Management (MDM)

## Naming Standards

# Security

# Appendix A – Acronyms List

|  |  |
| --- | --- |
| **Acronym** | **Definition** |
| HDFS | Hadoop Distributed File System |
|  |  |
|  |  |

# Appendix B – Tools Needed

Web Browser

Visio

Lync/Skype for Business/HipChat

UltraEdit/Ultra Compare

SublimeText 3

DQ Analyzer

Aqua Data Studio

Git

Cgywin (maybe)

PuTTY

FileZilla or WinSCP

Python

Tableau

VM (Oracle VM VirtualBox)

Admin Rights

Ability to install Linux (non-VM)