Slide 2

Kylo Deployment Fact and Data Lake Design

Le Nhu Chu Hiep

1. Kylo Overview

2. New Architecture

Section 1

Kylo Overview

What is Kylo ?

- A data lake management software leverage the power of apache nifi to manage the big data. It run on top of apache hadoop eco-system to satisfy the data lake constraints.
- Kylo provides a friendly Web application interface for user to create ingest schema, monitor data, security authentication and some limited analytic tool which run on top of apache spark.

Kylo Issues

- Lack of a setup support and detail document leading to many complex task during kylo installation.

Kylo Issues

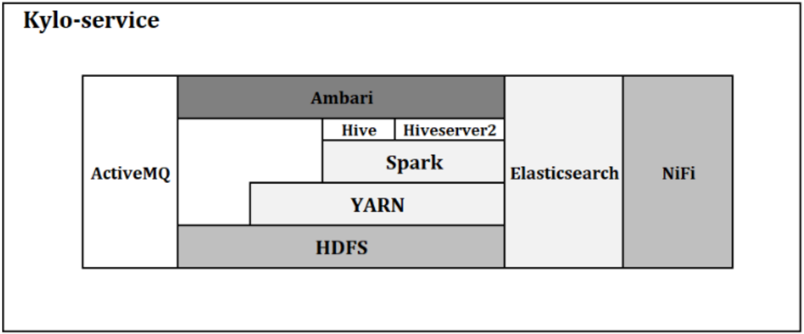
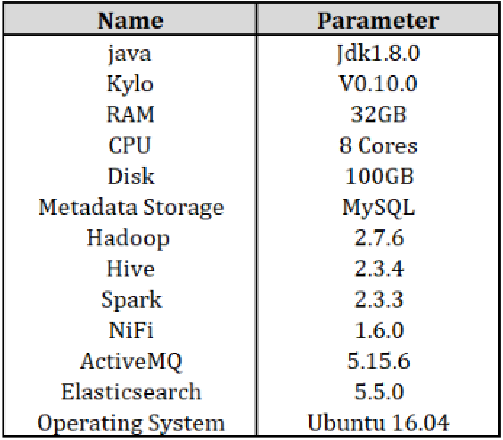


Figure 10. Basic Underlying Technologies Architectures

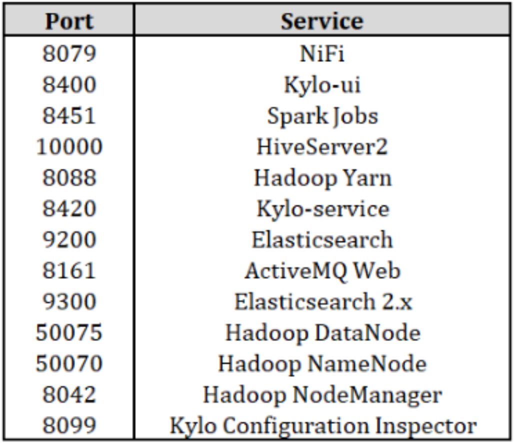
Kylo Issuses



Name	Parameter
java	Jdk1.8.0
Kylo	V0.10.0
RAM	32GB
CPU	8 Cores
Disk	100GB
Metadata Storage	MySQL
Hadoop	2.7.6
Hive	2.3.4
Spark	2.3.3
NiFi	1.6.0
ActiveMQ	5.15.6
Elasticsearch	5.5.0
Operating System	Ubuntu 16.04

Figure 12. Hardware Parameters & Software Version

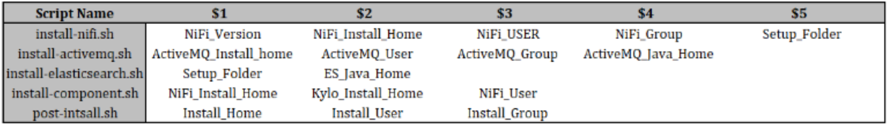
Kylo Issues



Port	Service
8079	NiFi
8400	Kylo-ui
8451	Spark Jobs
10000	HiveServer2
8088	Hadoop Yarn
8420	Kylo-service
9200	Elasticsearch
8161	ActiveMQ Web
9300	Elasticsearch 2.x
50075	Hadoop DataNode
50070	Hadoop NameNode
8042	Hadoop NodeManager
8099	Kylo Configuration Inspector

Figure 13. Network Ports List

Kylo Issues



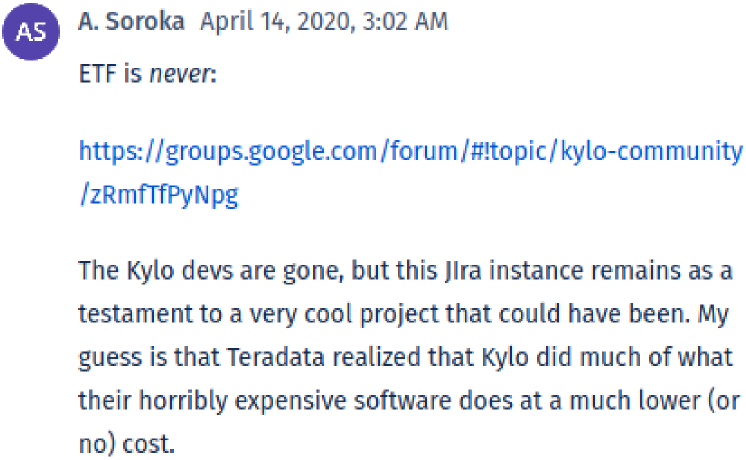
Script Name	$1	$2	$3	$4	$5
install-nifi.sh	NiFi_Version	NiFi_Install_Home	NiFi_USER	NiFi_Group	Setup_Folder
install-activemq.sh	ActiveMQ_Install_home	ActiveMQ_User	ActiveMQ_Group	ActiveMQ_Java_Home	
install-elasticsearch.sh	Setup_Folder	ES_Java_Home			
install-component.sh	NiFi_Install_Home	Kylo_Install_Home	NiFi_User		
post-intsall.sh	Install_Home	Install_User	Install_Group		

Figure 16. Kylo installation scripts list and input parameters order

Kylo Issues

- RIP developer team

Kylo Issues



AS

A. Soroka April 14, 2020, 3:02 AM

ETF is *never*:

https://groups.google.com/forum/#!topic/kylo-community/zRmfTfPyNpg

The Kylo devs are gone, but this JIra instance remains as a testament to a very cool project that could have been. My guess is that Teradata realized that Kylo did much of what their horribly expensive software does at a much lower (or no) cost.

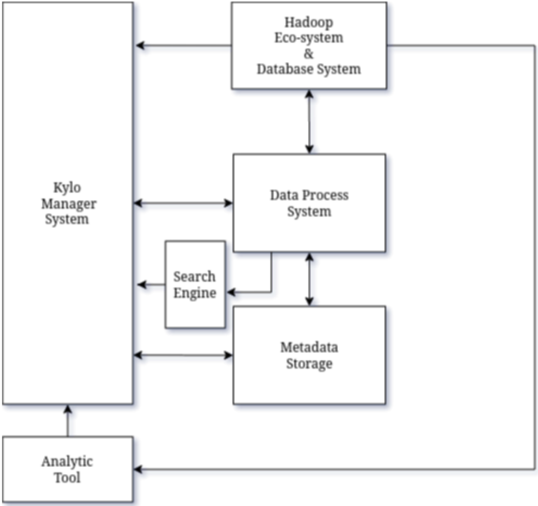
Kylo Issues

- Design limitation:
  - Structured vs Semi-structured only (spark SQL)
  - Limit in file size (spark SQL)
  - Not support multiple file ingest at once (apache Nifi)

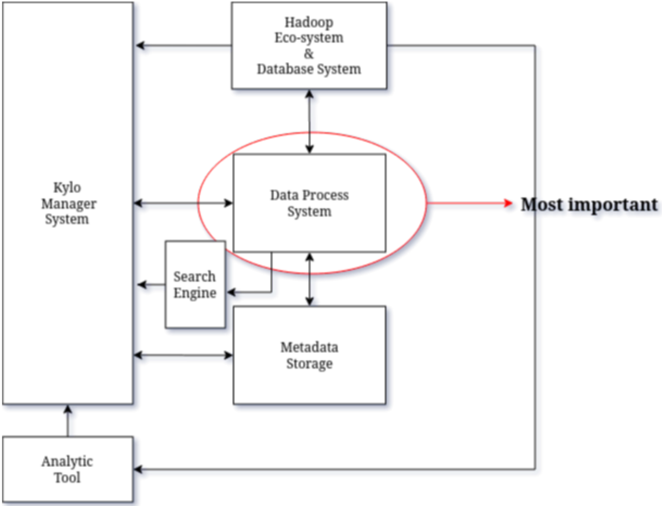
Section 2

New Architecture

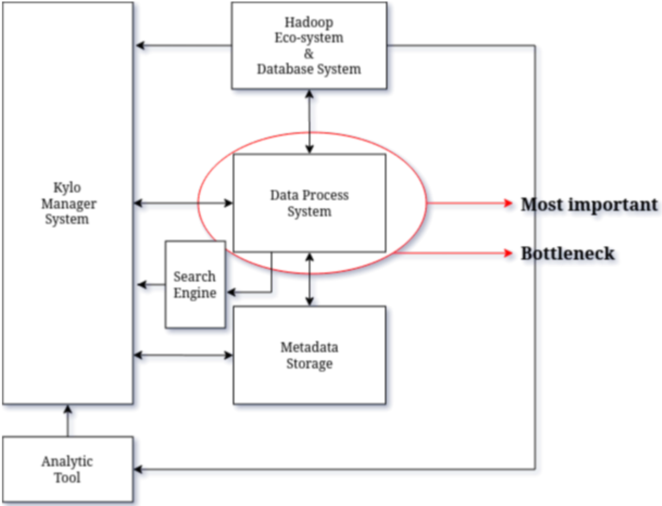
Kylo simplify architect



Kylo heart of system



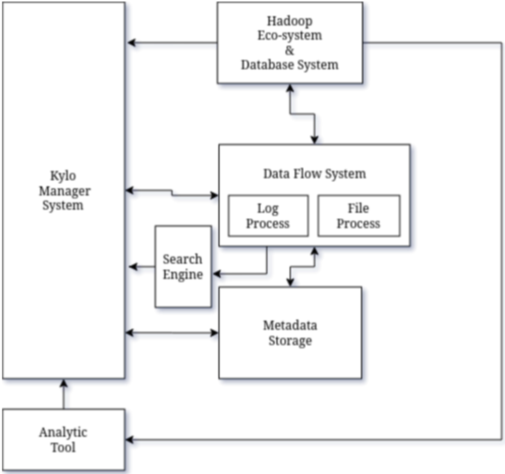
Kylo problem



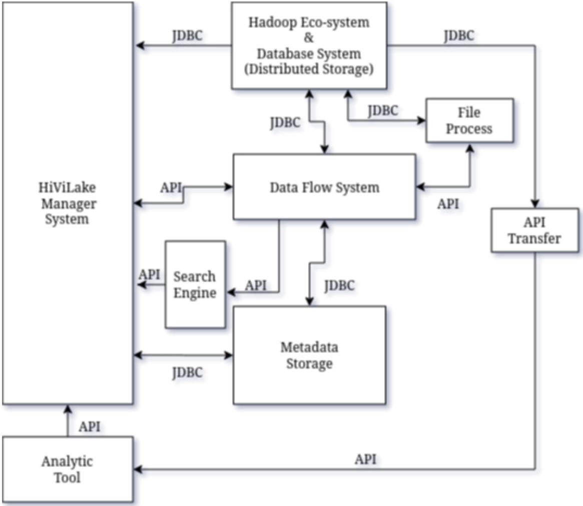
Kylo solution

Since the number of file ingestion in apache nifi cause problem, the solution should be seperate the file ingest and log ingest into 2 process, the nifi keeps track, audites and profiles data through log and the ingest of data will be handled be another services.

Kylo solution



HiViLake architect


Hadoop Eco-system
&
Database System
(Distributed Storage)
JDBC
JDBC
JDBC
JDBC
File
Process
HiViLake
Manager
System
API
Data Flow System
API
API
Transfer
API
Search
Engine
API
JDBC
Metadata
Storage
JDBC
API
Analytic
Tool
API