Big Data / ETL Automation Testing & Robot Framework

Robin Li

robinli@live.ca

Agenda

1. Big data Brief Introduction
2. ETL Automation Test
3. Robot Framework
4. Python Language
5. Q &A

Data --> Big Data

Why: BI / AI

Data is resource, ...

Quantity: KB(kilo), MB(Mega), GB(Giga), TB(Tera), PB(Peta),EB(Exa),ZB(Zetta),YB(Yotta)

Quality: QA / Testing

Data Base -> Data Warehouse -> Data Mart -> Data Lake

Challenges : Volume, Variety, and Velocity

Big Data Technology

Apache Hadoop：

Storage: HDFS + Computing: MapReduce

Main Platforms: Cloudera, Horonworks, ...

Tech: MPP Massive Parallel Processing

Some popular tools:

Data Management: HDFS, YARN

Operations: Zookeeper, Cloudbreak, Oozie

Data Access: Pig, Hive, Storm, Hbase, Spark,

Integration: Falcon, WebHDFS, Sqoop,Kafka

E T L / Data Flow

1. Data sources (SQL, GFF, CSV,PSV....)

Ingestion Parser/Mapper

2. Hadoop HDFS

Parquet file -- column-oriented

(Apache HIVE, Cloudera Impala, Pig)

Configuration files

3. Stream out for applications

Data Lake

Ingestion

Ingestion Architecture:

- Scalable, Extensible to capture streaming and batch data.
- Provide capability to business logic, filters, validation, data quality, routing, etc. business requirements.

Technology Stack:

- Apache Flume
- Apache Kafka
- Apache Storm
- Apache Sqoop
- NFS Gateway

Storage/Retention

Data Storage:

- Depending on the requirements data is placed into Hadoop HDFS, Hive, Hbase, Elastic Search or In-memory.
- Metadata management
- Policy-based Data Retention is provided.

Technology Stack:

- HDFS
- Hive Tables
- Hbase/MapR DB
- Elastic Search

Processing

Data Processing:

- Processing is provided for both batch and near-realtime use cases
- Provision Workflows for repeatable Data processing
- Provide Late Data Arrival Handling

Technology Stack:

- Map Reduce
- Hive
- Spark
- Storm
- Drill

Access

Visualization and APIs:

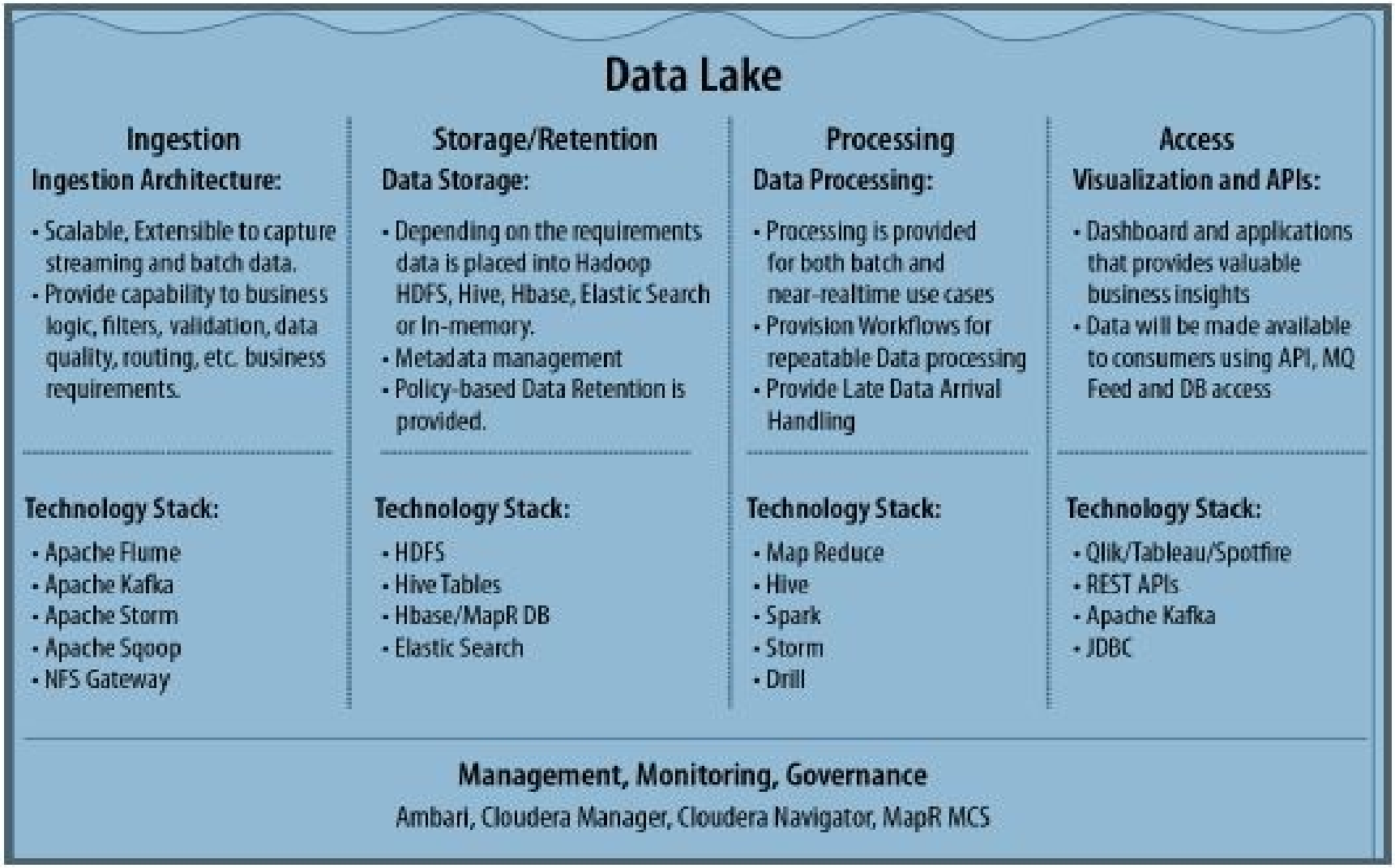
- Dashboard and applications that provides valuable business insights
- Data will be made available to consumers using API, MQ Feed and DB access

Technology Stack:

- Qlik/Tableau/Spotfire
- REST APIs
- Apache Kafka
- JDBC

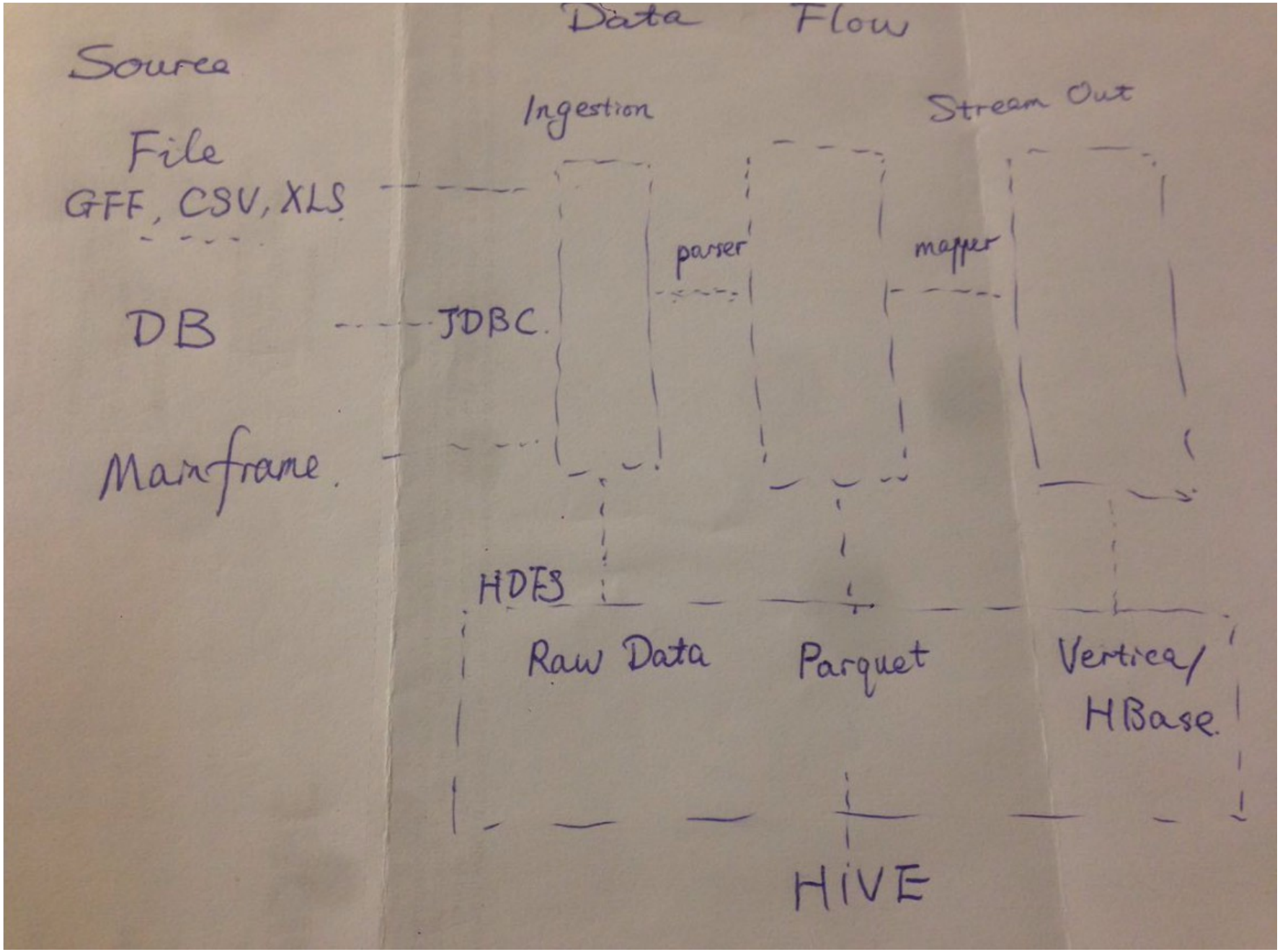
Management, Monitoring, Governance

Ambari, Cloudera Manager, Cloudera Navigator, MapR MCS



From Architecting Data Lakes

Data Management Architectures for Advanced Business Use Cases


Data Flow
Source
Ingestion
Stream Out
File
GFF, CSV, XLS
parser
mapper
DB
JDBC.
Mainframe.
HDFS
Raw Data
Parquet
Vertica/
HBase.
HIVE

Data Quality

1. Source quality – Extract Transform Load

   Bulk history: Sampling by order

   Delta data: Fully verification,

   Timestamp

2. Data Quality:

Integrity, Completeness, ...

(Counts, Null rate, MD5,Orphan records,log check.....)

Robot Framework

* Generic test automation framework
* ATDD: acceptance test-driven development
* Keyword-driven testing framework that uses tabular test data syntax

GUI: RIDE / CLI: pybot testSuite.txt ["Test Case"]

Connect to Test Case Management tools, pass test results to ALM /Jira

Devops: Build tools(Jenkins) + Deploy tools(UCD)

https://github.com/robotframework/QuickStartGuide/blob/master/QuickStart.rst

RIDE
(Robot framwork IDE)

1. Python 2.6 above --- Python 2.7.13
2. wxPython 2.8.12.1 with Unicode support
3. Robot framework
4. RIDE

https://github.com/robotframework/RIDE/wiki/How-To#starting-ride

Libraries

http://robotframework.org

Libraries

Standard: Builtin, Collections, Strings

External: Selenium, Database

In-house library/function

Example1: sql count(*) [(12,)] HIVE

2: “,” in csv file

Python Language

Easy, flexible Scripting Languages

Indent: 4 spaces to define block

Convention:

- Constant --UPPER
- camel case
- __function__, _function_, self.

official vs irregular

print-out vs hand-writing

Example :github.com/robin3795 sudoku game

Q & A

THANK YOU !