Hadoop - storage/processing

HDFS : block, replication 3,

processing : MR --- job tracker, task t

YARN -- resource , scheduler, application manager, node manager, application master

---MR, spark, TEZ

----------------------

data ingest

data pre-processing

data analysing

data processing

data store/export/report

data reporting

ooozie - scheduling tool

ingest - sqoop, flume, kafka

sqoop - RDBMS /strcutured -- full ingest, incremental ingestion - append/last mod

flume, kafka -semi strcutured, unstructured

data preprocessing - pig, programming

analysing - HIVE ----------> impala (intermediate operation cache)

hdfs files

create table t1 (id int, name string) location '\''

t1 t2 -- tables

hive processing -- select t1 join t2 (10 sec)

processing - hive (MR), Tz, spark --->

spark - in memory java scala python

spark - core ,spark sql---> tables (1sec)

files in HDFS

hive t3

HDFS --> sqoop export RDBMS

tables --hive impala tableau qlickview looker

structured data

sqoop---->hdfs-->preprocess-->hive tables/files---> preprocessing---> processing hive/spark---> store/export/visualize (fork)

----> preprocessing---> processing hive/spark---> store/export/visualize

ooozie

action

workflow (action scheduledc1

coordinator (

bundle

c1 c2 --> bundle

c1 c3 --> bundle

HBASE ---> NoSQL (storage hDFS)

Row based

noSQL ---> column based

c1, c2, c3

1,2,3,

1,2,3,4,5

dev / real time edge node - no UI access

quick start -- edge node UI

Hue --- cluster UI ---> quick start, windows URL--->

hbase --- > fast retrieval, hadoop be

c1,c2,c3,c4

1

1 1 1 1

1 1