Applications setups, commands, Project passages

Git

https://github.com/git-for-windows/git/releases/download/v2.41.0.windows.2/Git-2.41.0.2-64-bit.exe

AWS CLI

https://awscli.amazonaws.com/AWSCLIV2.msi

MAC AWS CLI

https://awscli.amazonaws.com/AWSCLIV2.pkg

NIfi download Link

https://archive.apache.org/dist/nifi/1.6.0/nifi-1.6.0-bin.zip

Cassandra Download

https://36lbuck.s3.amazonaws.com/datastax-community-64bit_2.2.3.msi?X-Amz-Algorithm=AWS4-HMAC-SHA256&X-Amz-Credential=AKIAT5PUAWQ7FI7G5YCH%2F20230819%2Fap-south-1%2Fs3%2Faws4_request&X-Amz-Date=20230819T025923Z&X-Amz-Expires=604000&X-Amz-SignedHeaders=host&X-Amz-Signature=821a349b4c988b6f184bac6158f9a4d62762a6b9191d9fbdcdd8c6dfe0d12970

WINSCP LINK TO DOWNLOAD

https://winscp.net/download/WinSCP-6.1.1-Setup.exe

Filezilla for Mac

https://dl3.cdn.filezilla-project.org/client/FileZilla_3.65.0_macosx-x86.app.tar.bz2?h=ibsShuwvqaG7liI4mG0InQ&x=1690095964

https://spark.apache.org/docs/2.2.0/streaming-kafka-0-10-integration.html

Find the second highest salary Reference

https://stackoverflow.com/questions/58490229/second-highest-value-by-department-using-apache-spark-dataframe

Winutils Setup

===

Goto any drive C or D drive
Create a folder D:/hadoop/bin
Paste the downloaded winutils file in **bin** folder

=================

AWS CLI MAC CURL COMMANDS
Works with this as well.

```
$ curl "https://s3.amazonaws.com/aws-cli/awscli-bundle.zip" -o "awscli-bundle.zip"
unzip awscli-bundle.zip
sudo ./awscli-bundle/install -i /usr/local/aws -b /usr/local/bin/aws
```

================

AWS Configure

=============

```
aws configure

AWS Access Key ID = AKIAS3H27Y6U3W4RLXNI
AWS Secret Access Key = 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
Default region name =  ap-south-1
Default output format = json
aws s3 ls
```

=================

Windows cmd Folks

```
notepad.exe zeyofile
dir
aws s3 cp zeyofile s3://36buck/URNAMEdir/
aws s3 ls s3://36buck/URNAMEdir/
```

AWS S3 Commands

==============================

```
aws s3 ls
aws s3 mb s3://URNAME36buck  (mb = make bucket)
aws s3 ls
aws s3 rb s3://URNAME36buck   (rb = remove bucket)
aws s3 ls
```

```
mkdir localdir
echo zeyo1> localdir/file1
echo zeyo2> localdir/file2
aws s3 sync localdir/  s3://buck36/36dir/URNAMEdir/  (sync is to upload file automatically)
aws s3 ls s3://buck36/36dir/URNAMEdir/
```

================

Linux/Mac

```
touch zeyofile
ls
aws s3 cp zeyofile s3://36buck/URNAMEdir/
aws s3 ls s3://36buck/URNAMEdir/
```

Windows cmd

```
mkdir localdir
cd localdir
notepad.exe file1
cd ..
aws s3 sync localdir/  s3://buck36/36dir/URNAMEdir/
```

Windows Users Play with S3

===================================

Step 1 -- install aws cli and Git bash windows users

Step 2 -- Open windows cmd / git bash

```
aws configure

AWS Access Key ID = AKIAS3H27Y6U3W4RLXNI
AWS Secret Access Key = 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
Default region name =  ap-south-1
Default output format = json
aws s3 ls
```

step 3 --

Git bash

```
mkdir localdir
echo zeyo1> localdir/file1
echo zeyo2> localdir/file2
aws s3 sync localdir/  s3://buck36/36dir/URNAMEdir/
aws s3 ls s3://buck36/36dir/URNAMEdir/
```

AWS EMR Delpoyment

==

*Step 1 --- PUT SOME UNIQUE in under <URNAME> AND EXECUTE Command

==

```
aws emr create-cluster --applications Name=Hadoop Name=Spark --ec2-attributes
'{"InstanceProfile":"EMR_EC2_DefaultRole","SubnetId":"subnet-
0440ec7d6b647d7c8","EmrManagedSlaveSecurityGroup":"sg-
024600bb0add188dd","EmrManagedMasterSecurityGroup":"sg-0bc726349e193f572"}' --
release-label emr-5.36.1 --log-uri 's3n://aws-logs-195947382697-ap-south-
1/elasticmapreduce/' --steps '[{"Args":["spark-submit","--deploy-mode","client","--
master","local[*]","--packages","org.apache.spark:spark-avro_2.11:2.4.7","--
class","pack.obj","s3://azeyo.dev/SparkDeploy-0.0.1-
SNAPSHOT.jar","<URNAME>dir"],"Type":"CUSTOM_JAR","ActionOnFailure":"CONTINUE
","Jar":"command-runner.jar","Properties":"","Name":"Spark application"}]' --instance-
groups'[{"InstanceCount":1,"EbsConfiguration":{"EbsBlockDeviceConfigs":[{"VolumeSpecifi
cation":{"SizeInGB":32,"VolumeType":"gp2"},"VolumesPerInstance":2}]},"InstanceGroupTyp
e":"MASTER","InstanceType":"r5.xlarge","Name":"Master Instance Group"}]' --
configurations '[{"Classification":"spark","Properties":{}}]' --auto-terminate --service-role
EMR_DefaultRole --name '<URNAME>Cluster' --scale-down-behavior
TERMINATE_AT_TASK_COMPLETION --region ap-south-1
```

==

Step 2 --- Note down the cluster id- Check the status of the Cluster id

==

```
j-1X7DXRL02Z3LD

aws emr describe-cluster --cluster-id <j-1X7DXRL02Z3LD> | grep 'State'
```

==

Step 3 --- Once your see terminating State-- check the target location

==

```
aws s3 ls s3://azeyo.dev/dest/
aws emr list-clusters --active | grep 'STARTING'
```

===

Nifi Installation in Windows

==
Download Nifi and Extract using 7z or WinRar
Always have nifi in drive not in a folder
Go inside nifi folder
Go inside bin folder
double click run-nifi file
It opens a command prompt

If any pop up comes and gets closed
Go inside nifi folder
Go to conf folder
Open nifi.properties
Change port from 8080 to 9090
Save and close
Again go to nifi folder --> bin folder --> double click run-nifi

JAVA_HOME can be one possible issue

Mac Folks

Go to Nifi extracted Folder using terminal
Go inside bin
trigger
sh nifi.sh start

Go to browswer (Same MAC or Windows)

hit the below URL
localhost:8080/nifi
localhost:9090/nifi
=====================

===

Commands to start ZooKeeper and Kafka

==

**Before we start kafka we should start zookeeper

```
-----------------------------
zoopkeeper start command
-----------------------------
cd C:\Windows\System32\cmd.exe
.\zkserver
```

```
-----------------------------
kafka start command
-----------------------------
cd D:\kafka_2.11
.\bin\windows\kafka-server-start.bat .\config\server.properties
```

```
-----------------------------
kafka topic create command (create new topic if needed -> change the topic in last [manipur1])
-----------------------------
cd D:\kafka_2.11\bin\windows
kafka-topics.bat --create --zookeeper localhost:2181 --replication-factor 1 --partitions 1 --topic manipur1
```

```
-----------------------------
kafka producer start command (change the topic in last [manipur1])
-----------------------------
cd D:\kafka_2.11\bin\windows
kafka-console-producer.bat --broker-list localhost:9092 --topic manipur1
```

```
-----------------------------
kafka consumer start command (change the topic in last [manipur1])
-----------------------------
cd D:\kafka_2.11\bin\windows
kafka-console-consumer.bat --zookeeper localhost:2181 --topic manipur1
```

==

To Run Nifi

```
----------------------------
Goto -> D:\nifi-1.6.0\bin
Double click on the -> run-nifi
```

Steps to execute the kinesis Code

======================

1---Open cmd authenticate AWS

======================

```
        aws configure
        accesskey      -- AKIAT5PUAWQ7FI7G5YCH
        secretkey      -- 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
        region         -- ap-south-1
        outputformat    -- json
        aws s3 ls                                   -------------test this command
```

======================

2--- create a kinesis stream

======================

```
aws kinesis create-stream --stream-name <UNIQUE_STREAM_NAME> --shard-count 1
aws kinesis list-streams
```

==============

Push data Plan A

==============

```
aws kinesis put-record --stream-name <UNIQUE_STREAM_NAME> --partition-key 123 --data firstmessage
aws kinesis put-record --stream-name <UNIQUE_STREAM_NAME> --partition-key 123 --data secondmessage
aws kinesis put-record --stream-name <UNIQUE_STREAM_NAME> --partition-key 123 --data thirdmessage
```

==============

Push data Plan B – if upper message is not sent due to base64 decode error

==============

```
aws kinesis put-record --stream-name kinesismq --partition-key 123 --cli-binary-format raw-in-base64-out --data firstmessage
aws kinesis put-record --stream-name <UNIQUE_STREAM_NAME> --partition-key 123 --cli-binary-format raw-in-base64-out --data secondmessage
aws kinesis put-record --stream-name <UNIQUE_STREAM_NAME> --partition-key 123 --cli-binary-format raw-in-base64-out --data thirdmessage
```

4 ----- Create or use existing eclipse Project

5 ----- add Kinesis spark jars and paste the code .. Ensure you give unique group id and proper kinesis stream name

==

=============================

hive hbase integration

=============================

```
create external table hbasehive_<urname>(hrow string,hid string,hname string,hcountry
string) STORED BY 'org.apache.hadoop.hive.hbase.HBaseStorageHandler' with
serdeproperties ("hbase.columns.mapping" = ":key,zcf:id,zcf:name,zcf:country")
tblproperties("hbase.table.name"="37htab");
select * from hbasehive_<urname>;
```

url --- e.cloudxlab.com
username --- zeyobronstudent2845
password --- I8LFCRIZ

===

Athena --- Business Intelligence

===

Possible

Just Like Hive
We can create table on top of s3 locations
Query would be faster (Backend athena runs on Presto Engine)
You can create any number of tables and Insert to other tables
Partitions and Bucketing are possible

Not Possible

Customized UDF to Athena is hard
Creation of Managed Tables are not allowed -- Only External

Glance R and D of Athena

Design Architecture of Athena - Hive
Parquet,avro support is there or not ? Yes / No
What backend engine in Engine - Presto Engine

Kafka handson -- Windows Folks

==

1) Download zookeeper and Kafka
2) Place it in drive (NOT INSIDE ANY SUB FOLDERS)
3) Extract Both of them
4) In the same Drive (E or D or C) -- remove the tmp if you have
5) Go inside zookeeper folder and Go inside bin folder and open cmd
6) then trigger below command and do not close that window just minimize it

```
.\zkserver
```

7) then come back go inside kafka folder and open cmd
8) Trigger below command and do not close that window just minimize it

```
.\bin\windows\kafka-server-start.bat .\config\server.properties
```

9) then come back go inside kafka folder---> Bin folder ----> windows folder open cmd
10) Execute below command to create kafka topic

```
kafka-topics.bat --create --zookeeper localhost:2181 --replication-factor 1 --partitions 1 --topic manipur1
```

11) Come to the same windows folder and again open cmd and execute below producer command -

```
kafka-console-producer.bat --broker-list localhost:9092 --topic manipur1
```

12) Come to the same windows folder and again open cmd and execute below consumer command

```
kafka-console-consumer.bat --zookeeper localhost:2181 --topic manipur1
```

13) Start pushing the data atleast 10 messages --- check the consumer console to validate the data

===

Kafka spark streaming steps

==================================

1 ---- Start Nifi

2-------Start zookeeper after removing tmp folder

In E or D or C -- remove the tmp if you have
Go inside zookeeper folder and Go inside bin folder and open cmd
then trigger below command and do not close that window just minimize it

```
.\zkserver
```

3 ---- start kafka service and create topic

then come back go inside kafka folder and open cmd
Trigger below command and do not close that window just minimize it

```
.\bin\windows\kafka-server-start.bat .\config\server.properties
```

then come back go inside kafka folder---> Bin folder ----> windows folder open cmd
Execute below command to create kafka topic

```
kafka-topics.bat --create --zookeeper localhost:2181 --replication-factor 1 --
partitions 1 --topic newtk
```

4 --- Configure invokehttp with below URL

remote url - https://randomuser.me/api/0.8/?results=10

5 --- Configure putkafka

```
known brokers - localhost:9092
topic name    - newtk
clientname    - zeyo
```

Ensure under settings -- check mark -- success and Failure

6 -- Open consumer console for newtk

Come to the same windows folder and again open cmd and execute below consumer command

```
kafka-console-consumer.bat --zookeeper localhost:2181 --topic newtk
```

8 -- Start Nifi check the consumer console for data

==

Cassandra Installation

==============================

Step 1 ---
Install python (uninstall if you have any other version in add or remove programs)

Step 2 ---- *Windows Folks -- Download and install it* - Datastax
https://drive.google.com/file/d/1rIwlS-MJiq3cFWY-RbwnhxveMURr2Ym_/view?usp=sharing

Step 3 --- Windows or Mac or Ubuntu
Download Cassandra

Windows
Install Python 3.7
Click the Drive Link Download and Straight away install
Download Cassandra.zip and extract it

MAC /UBUNTU

Download Cassandra.zip and extract it

1 - Windows After installing datastax
Open cassandra cql shell directly

2 -- Extract cassandra - Go inside bin-- type cassandra and enter
Again open cassandra CQL shell

3 -- Ubuntu/ mac --- Extract Cassandra Go inside bin -- type .\cassandra and enter
Minimize it and open the cmd line for the same folder and type .\cqlsh

```
CREATE KEYSPACE zzdb WITH replication = {'class':'SimpleStrategy',
'replication_factor' : 1};

describe keyspaces;
use zzdb;
describe tables;

create table ztab(id int PRIMARY KEY,name text);
insert into ztab(id,name) values(1,'sai');
insert into ztab(id,name) values(2,'zeyo');
select * from ztab;
```

Windows

----

0 --- download datastax, python and cassandra.zip
1 ---- download and install Python (Ensure you check box the Path)
2 ---- Install Datastax community.msi
3 ---- Open cassandra shell (If its working stop here)

```
        cqlsh> describe keyspaces;
        describe keyspaces;
        use system_auth;
        describe tables;
        select * from roles;
```

MAC
Plan A

```
        brew install python
        pip install cql
        brew install cassandra
        type cqlsh
```

Plan B

Extract the downloaded cassandra from announcement group
apache-cassandra-3.11.13-bin.tar
Go Inside cassandra extracted Folder
Go inside Bin
Open terminal
Type cassandra and give enter -- dont close it please
Again open other cmd in the same bin folder type .\cql.sh

Cassandra keyspace creation

============================

```
        CREATE KEYSPACE b36 WITH replication = {'class':'SimpleStrategy',
        'replication_factor' : 1};
        use b36;
        create table ztab(id int PRIMARY KEY, name text);
        select * from ztab;
        insert into ztab (id,name) values (1,'zeyo');
        select * from ztab;
        insert into ztab(id,name) values (2,'analytics');
        select * from ztab;
        insert into ztab (id,name) values (2,'aditya');
        select * from ztab;
```

Start Cassandra and create a table in the sparkcassandra with one column value
----

```
CREATE KEYSPACE zeyok WITH replication = {'class':'SimpleStrategy',
'replication_factor' : 1};

use zeyok;
CREATE TABLE zeyotk(
  value text  PRIMARY KEY
  );
```

==

Kafka message check Task

============================

Start zookeeper
remove tmp folder
Start kafka service
create one topic modelcheck
Send three message (firstmessage,secondmessage,thirdmessage)
In eclipse add spark jars
Add kafka connector jars
trigger with group id earlyid and offsetreset earliest
Start the trigger (you will early message and upcoming message also)
Once tested --- change group id latestid and offsetrest latest
Start the trigger and push some data through CLI (u will see only new data not older data)

Project Passage

============================

This is __ My total years of exp and Relevant experience.

I got chance to work on different Big Data Stack. Like (Hadoop, hdfs, hive, spark, sqoop, AWS). Recently i started migrating project AWS.
Before I used to in the data ingestion team where we used RDBMS as a source and we sqoop the data to HDFS and we processed it using hive and write to HDFS as avro. We use avro because of schema evolution. We have multiple RDBMS table in which we run multiple Sqoop Jobs and do the processing.

Later for example (1 year ago). I started working with Data application team where I have spark rigorously. In the data application. We have so many WEB apis coming with complex json json with different data models we almost run 7 spark jobs for different use cases like Customer data cleansing, Prediction Model spark jobs with currency conversions and few of the spark jobs do joins with AVRO data which generated during data ingestion and we write data to different HDFS directories as per the requirement also with complex nested data generation. My business uses impala to do analytics on processed data.

In the recent times we started migrating the jobs to AWS. with services s3 , EMR for spark jobs,Athena for Business analytics and ec2 for scheduling.

We have a done POC on EMR step executions run those spark jobs using
EMR command Runner.

Interview -- AWS

======================================

We have data sources from Webapi powered with SSL and AWS S3 along with snowflake.

We run our jobs in AWS EMR. We consume the data from all the sources and do the necessary processing and finally write the data to 2 different destination -- s3,snowflake

We perform all the necessary DSL operations in spark. We almost run 10-11 steps in AWS EMR UNDER STEP Execution.

Deployment
===============================

We create our own cluster in the daily basis and do the development/implementation.and terminate the cluster by end of the day. And copy the necessary copy intermediate to s3 for next day use.

Once the implementation - we commit the code to GIT .

For the production deployment - We run the Jenkins Pipeline created by Devops Team which would enable the Jar in the production s3 bucket.

However we will create a step execution command runner Automation emr script and test in the dev environment and take it to the production.

We schedule the Job using Nifi Running in EC2 machine. Processor Name (Execute processor).

=========================

Creation of Profiles

Use Resumes Samples to Build your Resume
DONT DONT DONT Copy points from Samples
Master Document -- Get the Points from Master Document
Post your Resumes Today's Itself
Put the Points which you feel comfortable
Definitely Mention aws s3, EMR, EC2 Points
Do not mention any version of the Tools in resumes
Put your experience in the Descending order (Current company should be at the top)
Make your resume as Unique as Possible
Upload it immediately after review corrections
Start hearing the interview Audios
Whenever you get some time start listing all real time scenarios
I will give project Passage for you But you can make it Unique phrases
Start solving scenarios
HR Contacts ---- 20-30 HR Contacts a day -- call them through Forums. Shall I share my resume
NEVER GIVE UPPPPPPPPPPPPPPPPPPPP