# APACHE IMPALA - LAB

### **Group Members**

Taha Rushain – 05811 Muhammad Siraj - 10255 Maria Zafar – 19842 M Sarmad Ali - 19828

## Table of Contents

Overview	2
Selling points	2
Components of a Impala Architecture	2
Demo	3
Installation	3
Projec Setup	6
Execution	7

## Apache Impala – Lab

#### Overview

To overcome the slowness of Hive Queries, Cloudera offers a separate tool and that tool is what we call Impala. Impala queries run very faster than Hive Queries even after they are more or less same as Hive Queries. Impala is not built on MapReduce, Impala has its own execution engine.

#### Selling points

- High-performance, low-latency SQL queries
- To share databases and tables between both Impala and hive it integrates very well with the Hive Metastore.
- It is Compatible with HiveQL Syntax
- We can easily integrate with Hbase database system and Amazon Simple Storage System.

We can perform interactive, ad-hoc and batch queries together in the Hadoop system, by using Impala's MPP (M-P-P) style execution along with other Hadoop processing MapReduce frameworks.

For querying analytic data it offers new possibilities. In addition, to query this type of data we can use exploratory data analysis and data discovery techniques.



#### Components of a Impala Architecture

#### Demo

#### Installation

Step 1: Install Cloudera QuickStart Container using the following command

docker pull cloudera/quickstart:latest

(base) node@machine:~\$ docker pull cloudera/quickstart:latest latest: Pulling from cloudera/quickstart Image docker.io/cloudera/quickstart:latest uses outdated schemal manifest format. 1d00652ce734: Already exists Digest: sha256:f91bee4cdfa2c92ea3652929a22f729d4d13fc838b00f120e630f91c941acb63 Status: Image is up to date for cloudera/quickstart:latest docker.io/cloudera/quickstart:latest

Step 2: Download the dataset from Kaggle

We are using UK house pricing dataset that contains information on all registered property sales in England and Wales that are sold for full market value which can be downloaded from the following link

https://www.kaggle.com/hm-land-registry/uk-housing-prices-paid?select=price\_paid\_records.csv

Unzip the dataset and place the zip file along with the code files in a folder that would be used inside the container, in my case it is /home/node/Desktop/BDA

**Step 3:** Start the Docker image by executing the following command (Change the '/home/node/Desktop/BDA' to your folder location)

docker run --hostname=quickstart.cloudera --privileged=true -t -v /home/node/Desktop/BDA:/user/cloudera/shared -i -p 8888:8888 -p 7180:7180 cloudera/quickstart /usr/bin/docker-quickstart

#### Here

- /home/node/Desktop/BDA:/user/cloudera/shared: a mapping from the dataset folder on my Desktop to a shared folder within the cloudera docker (can now use the dataset within docker)
- 8888:8888: port 8888 of container is mapped to 8888 of my hosted OS (can now access hue on my Ubuntu through localhost:8888)
- 7180:7180: mapping for Cloudera manager, required to start the docker
- /usr/bin/docker-quickstart: starts the Cloudera docker
- --privileged=true: This is required for HBase, MySQL-backed Hive metastore, Hue, Oozie, Sentry, and Cloudera Manager
- -t : provide a bash shell after starting
- -i : we want to use the terminal

After the image is completely loaded, you will have a prompt inside the docker instance like this



Step 4: Execute the following command to start cloudera manager

/home/cloudera/cloudera-manager - express



Step 5: Execute the following commands (one by one) to sync your Docker time with NTP

service ntpd restart ntpdate -u 0.centos.pool.ntp.org hwclock -systohc service ntpd restart [root@quickstart /]# service ntpd restart Shutting down ntpd: [ 0K ] Starting ntpd: [ 0K ] [root@quickstart /]# ntpdate -u 0.centos.pool.ntp.org 22 May 23:26:11 ntpdate[13380]: adjust time server 162.159.200.1 offset 0.052902 sec [root@quickstart /]# hwclock --systohc [root@quickstart /]# service ntpd restart Shutting down ntpd: [ 0K ] Starting ntpd: [ 0K ] Starting ntpd: [ 0K ] **Step 6:** Open cloudera manager at the following url with cloudera as username and password http://localhost:7180/

**Step 7:** Press the down arrow button next to 'Cloudera Quickstart' title, press Start to start all cloudera service and select start on the prompt as well, as in the given screenshot



•				
cloudera manager clusters	- Hosts - Diagnostics	- Audits Charts - Administration	🛱 🚜 Search (Hotkey: /)	Support- cloudera -
Home	All Hosts Roles		30 minutes preceding May	ý 23, 2020, 1:31 AM UTC
Status All Health Issues Configuratio	Host Templates	mands		Add Cluster
	Parcels		Try Cloudera Enterprise Data H	Hub Edition for 60 Days
Cloudera QuickStart (CDH 5.)	7.0	harts	30m 1h 2h 6h 1	12h 1d 7d 30d 🗷 🗸

Step 8: After that, we need to install CDH5 parcel, go the following Hosts > Parcels in the top menu

#### Step 9: Download the CDH 5 parcel by clicking the Download button in the second row.

cloudera manager	Clusters - Hosts - Diagnostics -	Audits Charts - Administration -		🛱 🔏 Search (Hotkey: /) Support - cloudera -
Parcels				Parcel Usage Configuration Check for New Parcels
Location	Cloudera QuickStart			
Cloudera QuickStart	Parcel Name	Version	Status	Actions
Available Remotely	ACCUMULO	1.7.2-5.5.0.ACCUMUL05.5.0.p0.8	Available Remotely	Download
	CDH 5	5.7.0-1.cdh5.7.0.p0.45	Downloading 339.7 f	MIB/1.3 GIB Cancel
Filters				

#### Projec Setup

**Step 10:** We would use Hue to upload to interface with HDFS and upload our dataset. Open the following url and use username/password 'cloudera'

http://localhost:8888/

		(	
			Y
cloudera			

#### Step 11: Open File Browser, create a folder 'impala-input'

📢 🖵 🐐 Query Editors 🗸 🛛 Data Browsers 🗸 🛛 Workflows 🗸 Search	🖺 File Browser 🛛 Job Browser 🕫 cloudera 🗸 😧 🍽 🍽
About Hue Quick Start Configuration Server Logs	
Quick Start Wizard - Hue™ 3.9.0 - The Hadoop UI	
An LIP A Dury Editors - Data Browsen - Workflows - Search	Birlin Browner 103 Job Browner 105 clouders ∨ 10 100 m
File Browser	
Search for file name 🛛 🕸 Actions 🗸 Move to trash	⊕ Upload マ O New マ
	D File
∉ Home / user / cloudera 🖌	Directory

## Create Directory

Directory Name impala-input

Cancel	Create

 $\times$ 

#### **Step 12:** Now open the 'impala-input' folder in Hue and upload out dataset.

Ciuc M Query Editors + Data Browsers + Worknows + Search				File Browser	Sob Browser O <sub>6</sub> cloudera •	0 10 0
E File Browser						
Search for file name 🗘 Actions 🗸 🗴 Move to trash 👻					() Upload ~	O New ~
æ Home / user / cloudera / impala-input ≠					✓ History	會 Trash
🔷 🕴 Name	Size	0 User	Group     Group	Permissions	Date	\$
■ 2		cloudera	cloudera	drwxr-xr-x	May 23, 2020 10:28 AM	
		cloudera	cloudera	drwxr-xr-x	May 23, 2020 10:29 AM	
price_paid_records.csv	2.2 GB	cloudera	cloudera	-TW-TT	May 23, 2020 10:29 AM	

#### Execution

Step 13: Open Impala editor in Hue

HUE	😤 Query Editors 🛩	Data Browsers v Workflows v Search	File Browser	Job Browser	0🖁 cloudera 🛩	0	100	
🖨 Abou	It HI 🖗 Hive	nitgration Server Logs						
Quick S	tart S DB Query tart Pig Job Designer	p.g. The Hadoop UI						

Step 14: Execute the following query to create DB and refresh

H)UE 🖀 Query Editors 🗸	Data Browsers 👻 Workflows 👻 Search
Impala Query Editor	My Queries Saved Queries History
Assist Settings Session	
	<pre>create database uk_houses; 3 3</pre>
uk_nouses	Execute Save as Explain Format or create a New query

(H)UE 🕋	Query Editors 🗸	Data Browsers 👻 Workflows 🛩 Search
💡 Impala	Query Editor My	Queries Saved Queries History
Assist Settings	Session	
<ul> <li>&lt; Suk_houses</li> <li>Tables</li> <li>⊞ price_paid</li> </ul>	(1) Q 3	CREATE TABLE uk houses.price paid ( 'Transaction_unique_identifier' STRING, 'Price' STRING, Date of Transfer' STRING, 'Property Type' STRING, 'Duration' STRING, 'Duration' STRING, 'Dustrict' STRING, 'County' STRING, 'Co

**Step 15:** Execute the following query to create the table for our dataset

CREATE TABLE uk\_houses.price\_paid (

```
`Transaction_unique_identifier` STRING,
```

`Price` STRING,

`Date\_of\_Transfer` STRING,

`Property\_Type` STRING,

`Old\_New` STRING,

`Duration` STRING,

`Town\_City` TINYINT,

`District` STRING,

`County` STRING,

`PPDCategory\_Type` STRING,

`Record\_Status\_monthly\_file\_only` STRING

```
)
```

ROW FORMAT DELIMITED FIELDS TERMINATED BY ","

LOCATION "hdfs:///user/cloudera/impala-input/";

#### Step 16: Check the table sample

Himpala Query Editor My								E FB	e Browser 🛛 💷 Job B	rowser O <sub>5</sub> cloudera +	<b>v</b> ~
	Queries Saved Queri	es History									
ssist Settings Session											Ð
( Luk_houses Tables I) Q.D Improg_paid I defined Preview S	CREATE TABLE Transact Transact Date of Property Sample data Sample	LE uk houses.price Lion Unique identif Transfer STRING, / Type STRING, / STRING, / STRING, STRIN	paid ( fier' STRING, e_only' STRING FERMINATED BY * rra/impala-input	·/";							
ata sample for price_p	paid			-14						₫ v	fiew more
transaction unique identifier	FIFTC III	CTINETAL COST OF INCOMPANY	property type		<b>CITE 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1</b>	TOWD CITY	district	county	nndeatenony type	record status monthly	file only
Transaction_unique_identifier	Price	Date of Transfer	Property_type	Old/New	Duration	town_city	District	County	PPDCategory_Type	Record Status - monthly	file_only
Transaction_unique_identifier Transaction unique identifier (81B82214-7FBC-4129-9F6B-4956B4	Price 4A663AD) 25000	Date of Transfer	Property_type Property Type T	Old/New	Duration	NULL	district District OLDHAM	County County GREATER MANCHESTER	PPDCategory_type	Record Status - monthly fil	file_only
Transaction_unique_identifier Transaction unique identifier (81B82214-7FBC-4129-9F6B-4956B4 (8046EC72-1466-42D6-A753-4956BF	Price 4A663AD} 25000 F7CD8A2} 42500	Date of Transfer 1995-08-18 00:00 1995-08-09 00:00	Property_type Property Type T S	Old/New N	Duration F	NULL NULL	district District OLDHAM THURROCK	County County GREATER MANCHESTER THURROCK	PPDCategory Type A A	Record Status - monthly fil	file_only
transaction_unique_identifier Transaction unique identifier (81882214-7FBC-4129-9F6B-4956B-4 (8046EC72-1466-42D6-A753-4956BF (278D581A-58F3-4FCE-AF62-4956D	Price Price 4A663AD} 25000 F7CD8A2} 42500 087691E6} 45000	Date of Transfer 1995-08-18 00:00 1995-08-09 00:00 1995-06-30 00:00	Property_type Property Type T S T	Old/New N N N	Duration F F	NULL NULL NULL	district District OLDHAM THURROCK SEDGEMOOR	County County GREATER MANCHESTER THURROCK SOMERSET	PPDCategory Type A A A A	record_status_monthly_ Record Status - monthly fil A A	file_only le only
Transaction_unique_identifier Transaction unique identifier (81882214-7FBC-4129-9F6B-4956B4 (8046EC72-1466-42D6-A753-4956BF (278D581A-5BF3-4FCE-AF62-4956D (1D861C06-A416-4865-973C-4956D)	Price Price 4A663AD} 25000 F7CD8A2} 42500 087691E6} 45000 B12CD12} 43150	Date         or         transfer           Date of Transfer         1995-08-18 00:00         1995-08-09 00:00           1995-08-09 00:00         1995-08-30 00:00         1995-11-24 00:00	Property_type Property Type T S T T	Old/New N N N N	Duration F F F F	NULL NULL NULL NULL NULL	assinct District OLDHAM THURROCK SEGGEMOOR NORTH BEDFORDSHIRE	County County GREATER MANCHESTER THURROCK SOMERSET BEDFORDSHIRE	ppdcategory_type PPDCategory Type A A A A A A	Record_status_monthly_ Record Status - monthly fil A A A A	file_only
Transaction_unique_leentifier Transaction unique identifier (8188/2214-7FBC-4129-9F68-49568/ (8046EC72-1466-42D6-A753-49568/ (2780581A-5BF3-4FCE-AF52-49560) (10861C06-A416-4659-973-49566) (D08645FD-A815-43A6-A78A-4956E)	Price Price 4A663AD} 25000 F7CD8A2} 42500 087691E6} 45000 812CD12} 43150 E58F1874) 18899	date_or_transfer           Date of Transfer           1995-08-18 00:00           1995-08-09 00:00           1995-06-30 00:00           1995-11-24 00:00           1995-06-23 00:00	property_type Property Type T S T T S S	Old/New N N N N N N	Duration F F F F F F	NULL NULL NULL NULL NULL NULL	distinct District OLDHAM THURROCK SEDGEMOOR NORTH BEDPROFSHIRE LEEDS	county County GREATER MANCHESTER THURROCK SOMERSET BEDFORDSHIRE WEST YORKSHIRE	ppdcategory_type PPDCategory Type A A A A A A A A A A	record_status_monthly_ Record Status - monthly fil A A A A A	file_only
Transaction_unique_leentifier Transaction_unique identifier (81882214-76-4129-06-68-49568- (9046E-C72-1466-4206-A753-49568F (2780581A-5873-4FCE-AF62-49560 (10681-567-A815-4306-A753-49566 (005864457-A815-4306-A78A-49566 (89554163-2037-476A-9AA-42380C	Price Price 4A663AD} 25000 F7CD8A2} 42500 087691E6} 45000 812CD12} 43150 E58F1874) 18899 DD0AE5C} 81750	Date of Transfer           Date of Transfer           1995-08-18 00:00           1995-08-09 00:00           1995-06-30 00:00           1995-06-30 00:00           1995-06-23 00:00           1995-06-23 00:00           1995-06-23 00:00	property_type Property Type T S T T S S S	Old/New N N N N N N N	Duration F F F F F F F	NULL NULL NULL NULL NULL NULL	district District CLDHAM THURROCK SEDGEMOOR NORTH BEDFORDSHIRE LEEDS SALISBURY	county County GREATER MANCHESTER THURROCK SOMERSET BEDFORDSHIRE WEST YORKSHIRE WILTSHIRE	ppdcategory_type PPDCategory Type A A A A A A A A A A A A A A A A A A A	record_status_monthly_ Record Status - monthly fil A A A A A A A	file_only
Transaction, unique, Jeentiter Transaction, unique, Identiter (808221-F76-0129-0F8-45568) (8082621-F76-0129-0F8-45568) (80846572-1466-4206-4753-49568) (10864557-04954-4366-4206-429680) (D0845576-04954-436-4494-423880) (FEI950272-5794-5579-437-0-42388)	Price           Price           4A663AD)         25000           F7CD8A2)         42500           987691E6)         45000           812CD12)         43150           E58F1874)         18899           DD0AE5C)         81750           KC94546A)         56000	bate         or_transfer           Date of Transfer         1995-08-18 00:00           1995-08-09 00:00         1995-06-30 00:00           1995-06-30 00:00         1995-06-23 00:00           1995-06-23 00:00         1995-06-23 00:00           1995-06-19 00:00         1995-03-10 00:00	Property_type Property Type T S T T T S S S S	N N N N N N N N N N N	Duration F F F F F F F F F	NULL NULL NULL NULL NULL NULL NULL	distinct Oberict OLDHAM THURROCK SEDGEMOOR NORTH BEDFORDSHIRE LEEDS SALISBURY WEST OXFORDSHIRE	county County GREATER MANCHESTER THURROCK SOMERSET BEDFORDSHIRE WEST YORKSHIRE WILTSHIRE OXFORDSHIRE	ppdcategory_type PPDCategory Type A A A A A A A A A A A A A A A A A A A	record status monthly Record Status - monthly fil A A A A A A A A A A A	file_only le only
Transaction unique Jeentitier Transaction unique Jeentitier (B1882214.7FBC-4129-9F68-49568/ (B0465C72-1466-4206-4753-49568/ (2780581-58F14-FCE-AF62-49560) (D08645FD-A915-43A6-A7BA-49566) (D08645FD-A915-43A6-A7BA-49566) (FB195C27-E790-45F0-437A-42388) (FB195C27-E790-45F0-437A-42388)	Price           Price           4A663AD}           25000           PF7CD8A2)           42500           307691E6)           45000           B12CD12)           43150           E58F1874)           18899           DD0AESC)           81750           IC94546A)           56000           3EB2D4C1}           31000	Date         Of           Date         of           Transfer           Date         of           1995-08-18         00:00           1995-08-09         00:00           1995-08-30         00:00           1995-06-30         00:00           1995-06-30         00:00           1995-05-19         00:00           1995-03-10         00:00           1995-03-20         00:00	property_type Property Type T S T T T S S S S S S	Old/New N N N N N N N N N N N N	Duration F F F F F F F F F F F	NULL NULL NULL NULL NULL NULL NULL NULL	alsinet OLDHAM OLDHAM THURROCK SEDGEMOOR NORTH BEDPORDSHIRE LEEDS SALISBURY WEST OXFORDSHIRE RESTORMEL	county County GREATER MANCHESTER THURROCK SOMERSET BEDFORDSHIRE WILTSHIRE OXFORDSHIRE CORNWALL	ppdcategory_type PPDCategory Type A A A A A A A A A A A A A A A A A A A	Record Status - monthly fi A A A A A A A A A A A A A	file_only le only
Transaction, unique, Jeentifier Transaction, unique, Identifier (B188211-A756-4129-4768-49568) (B084EC72-1466-4206-4753-49568) (2780581-568-426-4753-49568) (108615-476-476-4956-4753-49569) (108615-476-476-476-476-476-476-476-476-476-476	Price           Price           4A663AD)         25000           F7CD6A2)         42500           087601E6)         45000           812CD12)         43150           558F1874)         18890           DD0AE5C)         81750           0C34546A)         56000           FRE12AFC1}         31000           FRE12AFC1}         82000	aate         on_transfer           Date         of Transfer           1995-08-18 00:00         1995-08-18 00:00           1995-08-20 00:00         1995-08-20 00:00           1995-08-20 00:00         1995-08-20 00:00           1995-08-19 00:00         1995-08-20 00:00           1995-08-10 00:00         1995-08-20 00:00           1995-08-20 00:00         1995-08-20 00:00	property.type Property Type T S T T T S S S S S S S	Old/New N N N N N N N N N N N N N	Duration F F F F F F F F F F F F F F	NULL NULL NULL NULL NULL NULL NULL NULL	district Olimitet OLIMAM THURROCK SEDGEMOOR SEDGEMOOR NORTH BEDPORDSHIRE LEEDS SALISBURY WEST OXFORDSHIRE RESTORMEL EALING	County County OREATER MANCHESTER THURROCK SOMERSET BEDFORDSHIRE WEST VORKSHIRE WILTSHIRE OXFORDSHIRE CORWALL OREATER LONDON	ppdcategory_type PPDCategory Type A A A A A A A A A A A A A A A A A A A	Record status - monthly II A A A A A A A A A A A A A A A A	file_only le only
Transaction, unique, joentimer Transaction, unique, identifier (8082214-756-4129-4768-49568/ (80826214-756-420-4753-49568/ (2705081-567-34FC5-4762-49560) (1D8615-06-416-4865-973C-49560) (006945FD-613-43A-678A-4956) (9054-563-2037-476A-9AA-423880) (006945FD-613-43A-678A-423880) (1D68015E1-74-553-4400-A2C1-42388) (000274-745-55-4400-A2C1-42388) (000274-745-55-4400-A2C1-42388)	price           4A663AD}         25000           F7CD8A2         42500           812CD12         43150           559F1874         18899           DD0AESC         81750           GC4454AA         56000           38B2D4C13         31000           SH2C6F4         82000           558268D4         10000	aate         org. transfer           Date of Transfer         11995-08-30 00:00           1995-08-30 00:00         1995-08-30 00:00           1995-08-30 00:00         1995-08-30 00:00           1995-08-23 00:00         1995-08-20 00:00           1995-08-30 00:00         1995-08-10 00:00           1995-08-10 00:00         1995-08-10 00:00           1995-08-10 00:00         1995-08-10 00:00	property.type Property Type T S T T T S S S S S S T T T T T T T T T T T T T	Old/New N N N N N N N N N N N N N N N	Uration F F F F F F F F F F F F F F F F	NULL NULL NULL NULL NULL NULL NULL NULL	distinct District OLDHAM THURROCK SEDGEMOOR NORTH BEDFORDSHIRE LEEDS SALISBURY WEST OXFORDSHIRE RESTORMEL EALING RHONDDA	county County OREATER MANCHESTER THURROCK SOMERSET BEDFORDSHIRE WILTSHIRE OXFORDSHIRE CORNORSHIRE CORNWALL GREATER LONDON MID GLAMORGAN	ppdcategory_type PPDCategoryType A A A A A A A A A A A A A A A A	record status monthly. Record Status - monthly fil A A A A A A A A A A A A A A A	file_only
Transaction unique Jeentimer Transaction unique Jeentimer (01882214.7FBC-4129.9F68-49566- (00486C72.1466.4206-4733.495687 (2780591A.86F3.4FCE-AF62.49660 (106864663-2037-4164-424960 (09564663-2037-478A-34A-42380 (09564663-2037-478A-34A-42380 (105601C-0C33-4147-8A21.42388 (1006017423-C33-4481-3424-3280 (00027423-C33-431-3484-32380) (00027423-C33-431-3484-32380)	price           Price           Ad663AD)         25000           F7CD8A2)         42500           987691E0)         45000           B12C012)         43150           SD6745E0)         81750           SD6745E0)         81750           EB12C012)         13100           BE822C012)         31000           BE822C012)         10000           B6825ABD         10500	aate         org         variation           Date         of Transfer           1999-08-18 00.00         1999-08-18 00.00           1999-08-18 00.00         1999-08-18 00.00           1999-08-18 00.00         1999-08-18 00.00           1999-08-18 00.00         1999-08-18 00.00           1999-08-10 00.00         1999-08-17 00.00           1999-08-17 00.00         1999-08-17 00.00	property.type Property.Type T S T S S S S S S S S S S S S S	Old/New N N N N N N N N N N N N N N N N N	Curation F F F F F F F F F F F F F F F F C C	NULL NULL NULL NULL NULL NULL NULL NULL	assuer Desinier OLDHAM THURROCK SEDGEMOOR NORTH BEDPORDSHIRE LEEDS SALISBURY WEST OXFORDSHIRE RESTORMEL EALING RHONDDA FYLDE	County GREATER MANCHESTER THURROCK SOMERSET BEEPORDSHIRE WEIST YORKSHIRE WILTSHIRE CORNWALL GREATER LONDON MID GLANORGAN LANCASHIRE	ppdcategory_type PPDCategoryType A A A A A A A A A A A A A A A A A	record status - monthy fi A A A A A A A A A A A A A A A A A A A	file_only le only

Step 17: Execute the following query to find the number of properties by Property Type

