

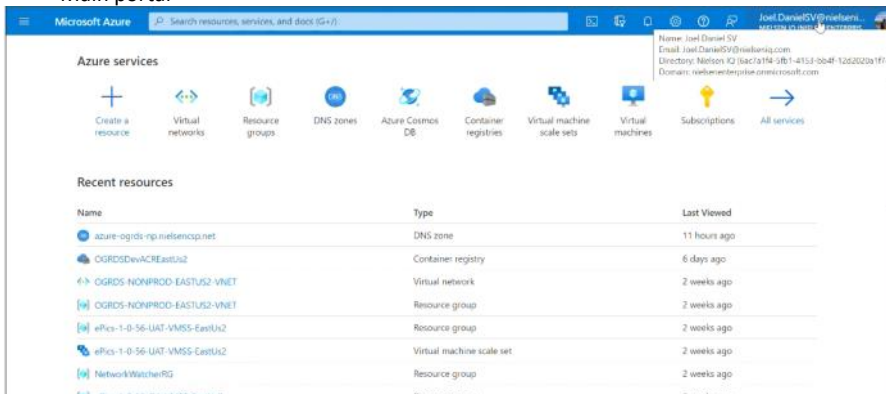
Cloud and Data Engineering

Monday, March 21, 2022 10:59 AM

- Better than on premise databases
- Less operational costs
- More security
- Example: GCP, Azure, AWS
- Azure offers 30+ services
- On premise resources requires more cost than cloud databases
- More scalability in cloud
- Cloud is cost effective.
- In our company we use azure cloud services.

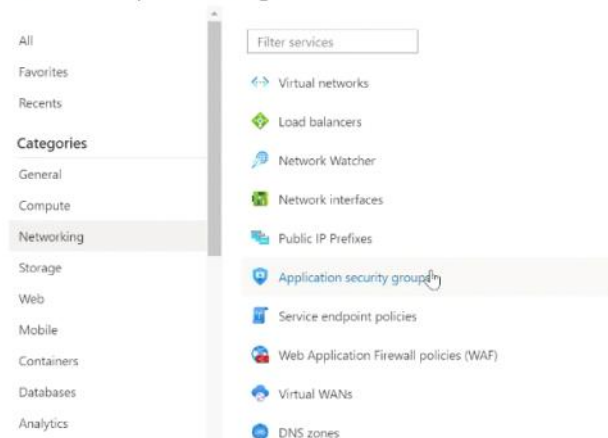
Services that we use

- Main portal

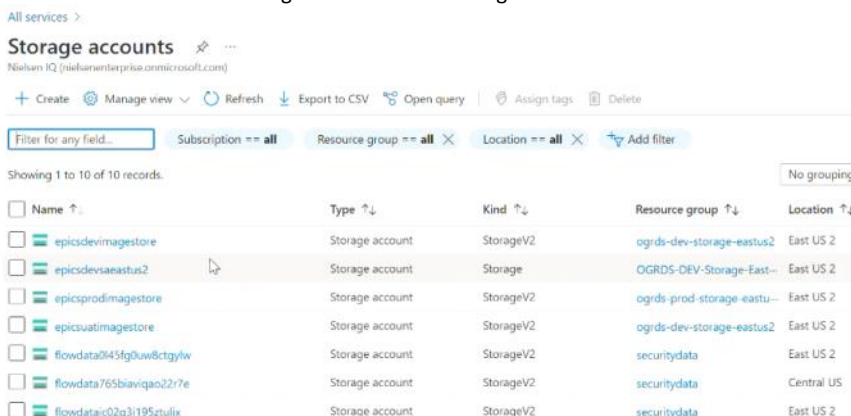


- Some of the Services provided by azure

All services | Networking



- What is DevOps: a person who do development and deployment is DevOps engineer
- We can store all the images and videos in storage accounts

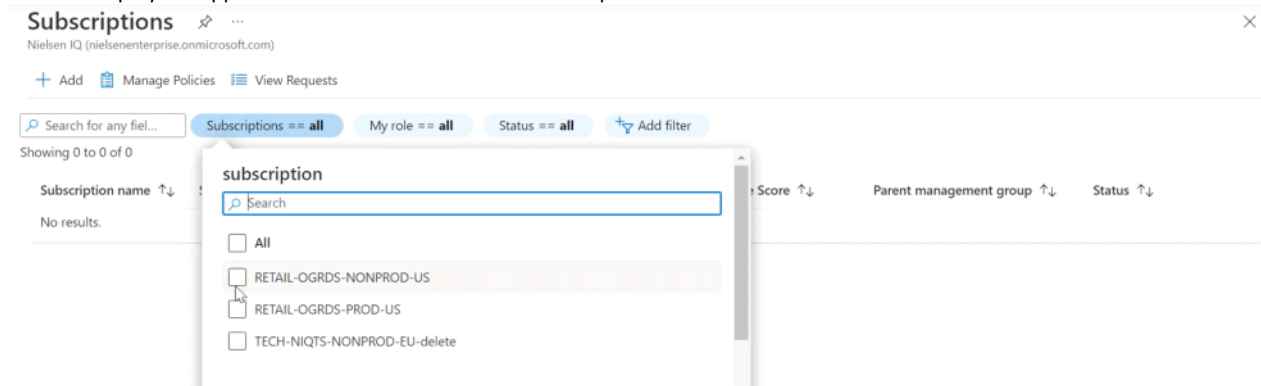


```

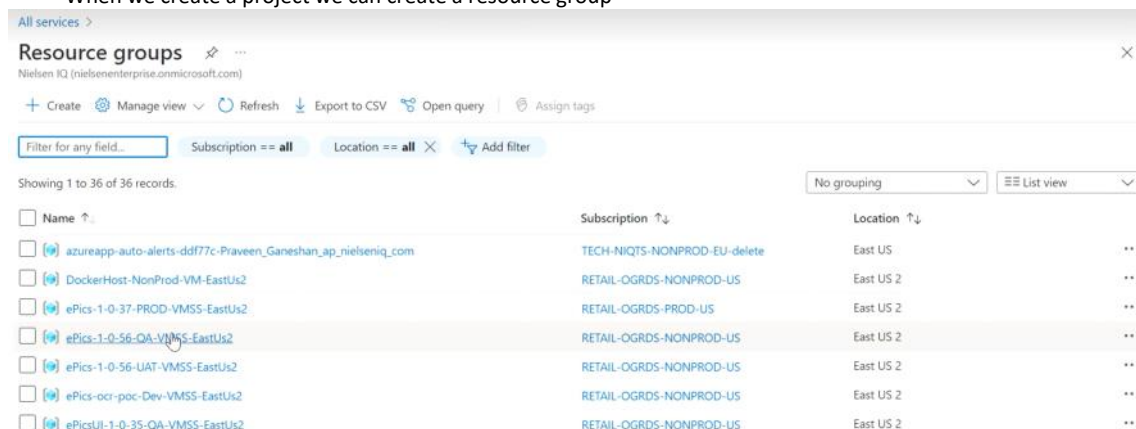
1 FROM maven:3.5-jdk-8-alpine AS build
2 WORKDIR /code
3
4 COPY pom.xml /code/pom.xml
5
6 # Adding source, compile and package into a fat jar
7 COPY ["src/main", "/code/src/main"]
8 RUN ["mvn", "clean", "install"]
9
10 FROM mcr.microsoft.com/java/jre:11u11-zulu-alpine
11
12 RUN apk add --update curl
13
14 # Get the Datadog APM tracing client
15 RUN curl --location --output /dd-java-agent.jar https://dtdg.co/latest-java-tracer
16
17 COPY --from=build /code/target/ePicsStorage-0.0.1-SNAPSHOT.jar /ePicsStorage.jar
18 CMD ["java", "-javaagent:/dd-java-agent.jar", "-XX:+UnlockExperimentalVMOptions", "-jar", "/ePicsStorage.jar"]

```

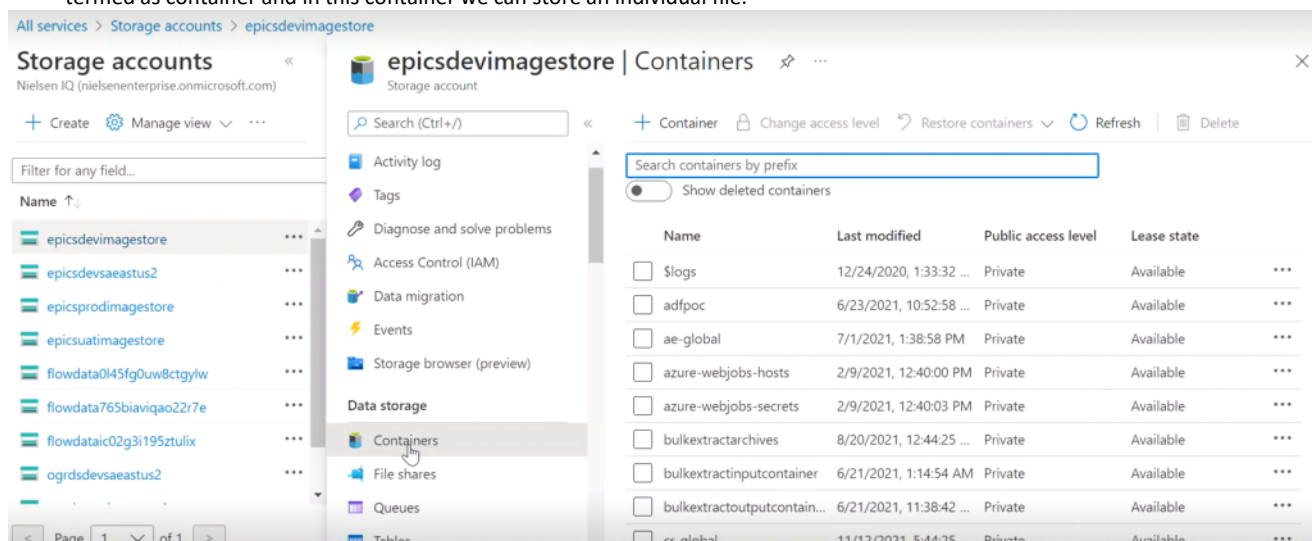
- To deploy the applications into cloud we use Azure Devops.



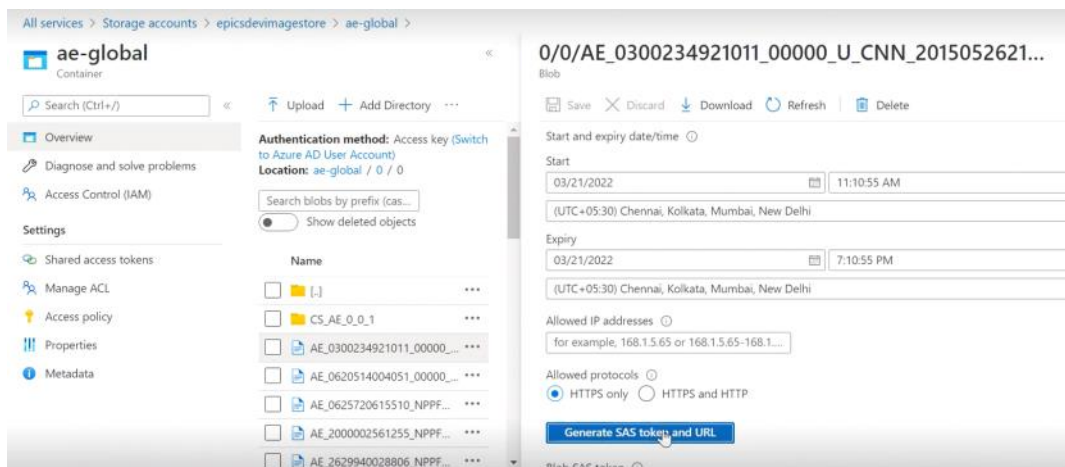
- To create resource in Azure we create a subscription using subscription service.
- When we create a project we can create a resource group



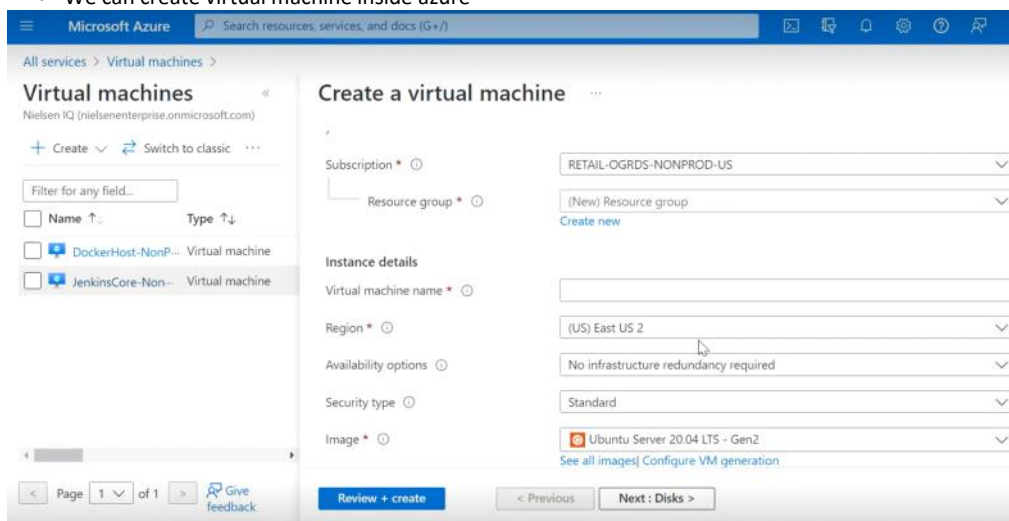
- Under resources we will have all the resources available for that project
- If we want to store any media in azure we can store it under storage accounts. Here folder is termed as container and in this container we can store an individual file.



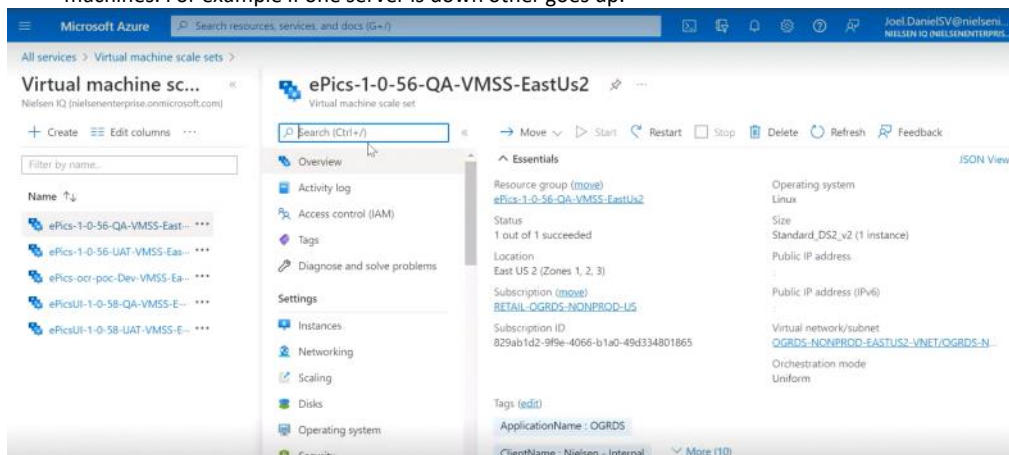
- Suppose if we want to give access or share the particular files to any one we can give them access through links by modifying permissions and generate a sas key as shown below. And we can use that sas key to share or utilize.



- We can access private images using sass key according to the permissions given.
- We can even utilize the azure sql services similar to on premise but we need to pass the given jdbc link in our project.
- We can create virtual machine inside azure



- Virtual machine scales set is used to manage the traffic or load balancing between virtual machines. For example if one server is down other goes up.



- We can create our own function to perform on any content or an image using function app, we can store functions and it's code on the cloud. So, the function performed on the image in the cloud

Microsoft Azure

Search resources, services, and docs (G+)

All services >

Function App

Nielsen IQ (nielsenenterprise.onmicrosoft.com)

+ Create Manage view Refresh Export to CSV Open query Assign tags Start Restart Stop Delete

Filter for any field... Subscription == RETAIL-OGRDS-PROD-US Resource group == all Location == all Add filter

Showing 1 to 2 of 2 records.

No grouping List v

Name	Status	Location	Pricing Tier	App Service Plan	Subscription
BulkExtractInputHandlerProd	Running	East US 2	Premium V2	premiumplanv2eastus2	RETAIL-OGRDS-PROD...
ThumbnailProd	Running	East US 2	Dynamic	EastUS2Plan	RETAIL-OGRDS-PROD...

- To trigger an even we need an event subscriptions for example if any "event" is occurred we call a "function"

Microsoft Azure

Search resources, services, and docs (G+)

All services >

Event Subscriptions

Event Grid

+ Event Subscription Refresh

Filter Filter with this... Topic Type Storage Accounts (Blob & GPv2) Subscription RETAIL-OGRDS-NONPROD-US Location East US 2

Name	Topic	Endpoint
ThumbnailGeneration	epicsdevimagestore	GenerateThumbnail
ThumbnailGenerationUAT	epicsuatimagestore	GenerateThumbnailUAT
eb6d8c25-3a82-c25c-78cd-1d253b57fa09	epicsdevimagestore	https://pmeastus2.svc.datafactory.azure.com:4443/triggerevent/BlobEventsTri...
96807567-d790-74c4-bd8d-86842d358d0d	epicsuatimagestore	https://pmeastus2.svc.datafactory.azure.com:4443/triggerevent/BlobEventsTri...

- The pricing also varies, premium means we need to pay even if we not use. And dynamic pricing means it depends on number of times we use the service.

Pricing Tier

Premium V2

Dynamic

- In key vaults we can use to store the confidential information

All services > Key vaults >

Key vaults

Nielsen IQ (nielsenenterprise.onmicrosoft.com)

+ Create

Filter for any field...

Name

- OGRDSKVVAUTOUS
- OGRDSKVQAUS
- OGRDSKVUATUS

Overview

Activity log

Access control (IAM)

Tags

Diagnose and solve problems

Events

Settings

Keys

Secrets

Certificates

Access policies

OGRDSKVQAUS

Key vault

Search (Ctrl+/)

Delete Move Refresh Open in mobile

Upcoming TLS 1.0, 1.1 deprecation: Please enable support for TLS 1.2 on clients (applications/platform) to avoid any service impact. Learn more here.

Essentials

Resource group (move) OGRDS-KeyVault-EastUs2-RG

Location East US 2

Subscription (move) RETAIL-OGRDS-NONPROD-US

Subscription ID 829ab1d2-9f9e-4066-b1a0-49d334801865

Vault URI https://ogrdskvqaus.vault.azure.net/

Sku (Pricing tier) Standard

Directory ID 6ac7a1f4-5fb1-4153-bb4f-12d2020a1f7d

Directory Name Nielsen IQ

Soft-delete Enabled

Purge protection Disabled

- Domain Name system

All services >

DNS zones

Nielsen IQ (nielsenenterprise.onmicrosoft.com)

+ Create Manage view Refresh Export to CSV Open query Assign tags

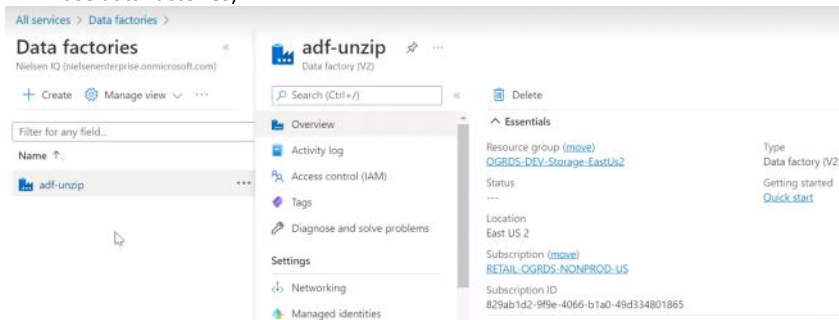
Filter for any field... Subscription == RETAIL-OGRDS-NONPROD-US Resource group == all Location == all Add filter

Showing 1 to 1 of 1 records.

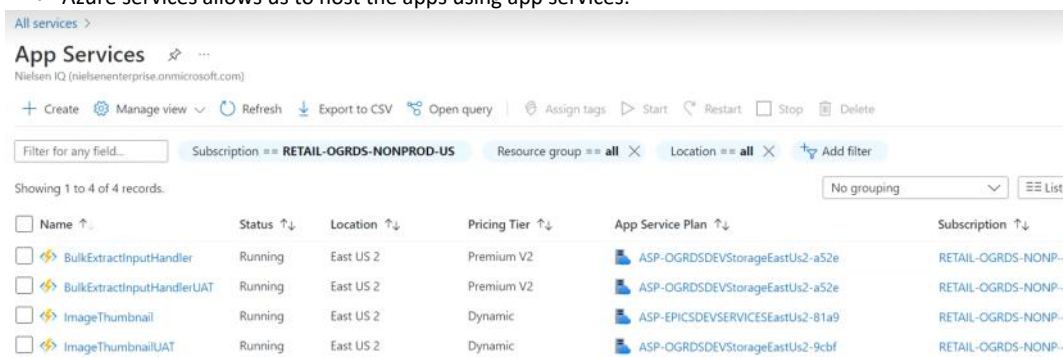
No grouping

Name	Num...	Resource group	Location
azure-ogrds-np.nielsenscp.net	19 / 10000	OGRDS-DEV-dns-EastUs2	Global

- Data Factories (If we want to move a large amount of database from a part to other part we use data factories)



- In epics , all that metadata is stored in mongo db collections.
- Meta data in mongo dB
- Image in Azure cloud storage accounts.
- And for those images we use function apps and generate thumbnail images.
- We upload the files from the user using spring boot and java , here java offers methods to upload an image into azure cloud.
- Load balance automatically creates a new virtual machine according to the loads that we defined as minimum and maximum.
- Azure services allows us to host the apps using app services.



Cloud Computing


Tuesday, March 22, 2022 11:24 AM

THE PROBLEM	
What ?	• On-Premises Architecture
Why ?	• Cost, Time & Effort ...
Where ?	• Installed Locally
When ?	• Cloud - 1990s started & 2000s – became mainstream (Emergence of Big 3)
Solution ?	• Cloud Computing

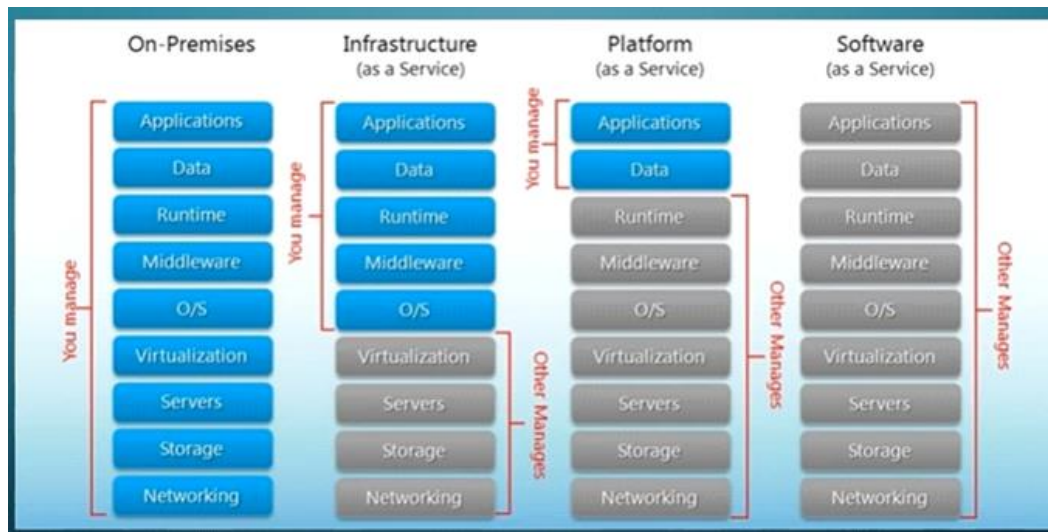
On Premise Architecture

	<ul style="list-style-type: none">• <u>Pros:</u><ul style="list-style-type: none">• Ownership/ complete control• No external factors/ Internet• <u>Cons:</u><ul style="list-style-type: none">• Cost/ Resources• Maintenance/ Upgrades
--	---

- If we want to store any secured information that should not be seen by any companies we store them on-premise.
- When we have any trouble or any update on the servers , if we are using cloud services. Cloud technicians will work on them and resolve the problem. And if we want to do the same thing locally we must maintain many number of experts to solve the issue. This results in high input cost.
- Enterprise edition of the cloud will have more amount of services by the cloud provider. Their response will be very fast.

SaaS	PaaS	IaaS
<ul style="list-style-type: none">▪ Software as a Service▪ SaaS platforms make software available to users over the internet, (usually for a monthly subscription fee) 	<ul style="list-style-type: none">▪ Platform as a Service▪ A PaaS vendor provides H/W & S/W over the internet to develop applications.▪ PaaS users tend to be developers.	<ul style="list-style-type: none">▪ Infra as a Service▪ IaaS businesses offer services such as pay-as-you-go storage, N/W, and virtualization.▪ IaaS gives users cloud-based alternatives to on-premise infrastructure.

- For IaaS we will have more control on the cloud than the SaaS.



AWS	MSA	GCP
<ul style="list-style-type: none"> Amazon Launched in 2006 Customers – Netflix, Samsung, MI, Airbnb 	<ul style="list-style-type: none"> Microsoft Launched in 2010 Customers – Honeywell, Apple, HP 	<ul style="list-style-type: none"> Google Launched in 2011 Customers – HSBC, PayPal, Dominos

- Azure Is leading the market , followed by AWS and GCP.
- Coding in cloud: Spark with Scala , Python , Snowflake (Datawarehouse) , For ETL purpose we use Airflow.
- Azure Active Development is used to manage the cost and resources among the groups that we have under our plan. It's similar to a boss managing it's resources to it's employee and restricting the un usage cost.
- It will be present under the groups section, where we can create a user and assign policies that he needs to follow.

The screenshot shows the Microsoft Azure portal interface. The top navigation bar includes the Microsoft Azure logo, a search bar, and various icons. The main content area displays the "Nielsen IQ | Overview" page, which is part of the Azure Active Directory service. The left sidebar contains a "Manage" section with links to Users, Groups, External Identities, Roles and administrators, Administrative units, Enterprise applications, Devices, App registrations, and Identity Governance. The main content area shows the "Overview" tab, which includes a search bar and a table of basic information.

Basic information			
Name	Nielsen IQ	Users	116,842
Tenant ID	6ac7a1f4-5fb1-4153-bb4f-12d2020a1f7d	Groups	104,316
Primary domain	nielsenenterprise.onmicrosoft.com	Applications	2,961
License	Azure AD Premium P2	Devices	53,855

- We can manage the virtual machines under VM's.
- The cost depends on the zones we selected, the less distance of the zone the less cost is.
- Under data bricks we can create clusters.
- Databricks is used for processing large amount of data.

- If we want to process big data use ADLS Gen 2
- We can store our files under Storage Accounts of Microsoft azure , we have separate UI for storage explorer.

BIG DATA

- BIG DATA means huge volume of the data



Big Data - Massive amount of data sets that cannot be stored, processed, or analyzed using traditional tools

Reason - Social media platforms and Internet

Example – FB – >500TB/Day – Pics, Videos, Messages ...

Data Types – Structured, Semi- Structured, Unstructured

Structured – CSV, Excel - Tables

Semi - Structured – XML, JSON - Emails

Unstructured – Audio, Video, Images - Fingerprints

Risk management - Banco de Oro, a Philippine banking company, uses Big Data analytics to identify fraudulent activities and discrepancies

Innovations - Rolls-Royce, one of the largest manufacturers of jet engines for airlines and armed forces across the globe, uses Big Data analytics to analyze how efficient the engine designs are and if there is any need for improvements.

Decision Making - Starbucks uses Big Data analytics to make strategic decisions. For example, the company leverages it to decide if a particular location would be suitable for a new outlet or not. They will analyze several different factors, such as population, demographics, accessibility of the location, and more.

Improve Customer Experience - Delta Air Lines uses Big Data analysis to improve customer experiences. They monitor tweets to find out their customers' experience regarding their journeys, delays, and so on. The airline identifies negative tweets and does what's necessary to remedy the situation. By publicly addressing these issues and offering solutions, it helps the airline build good customer relations.

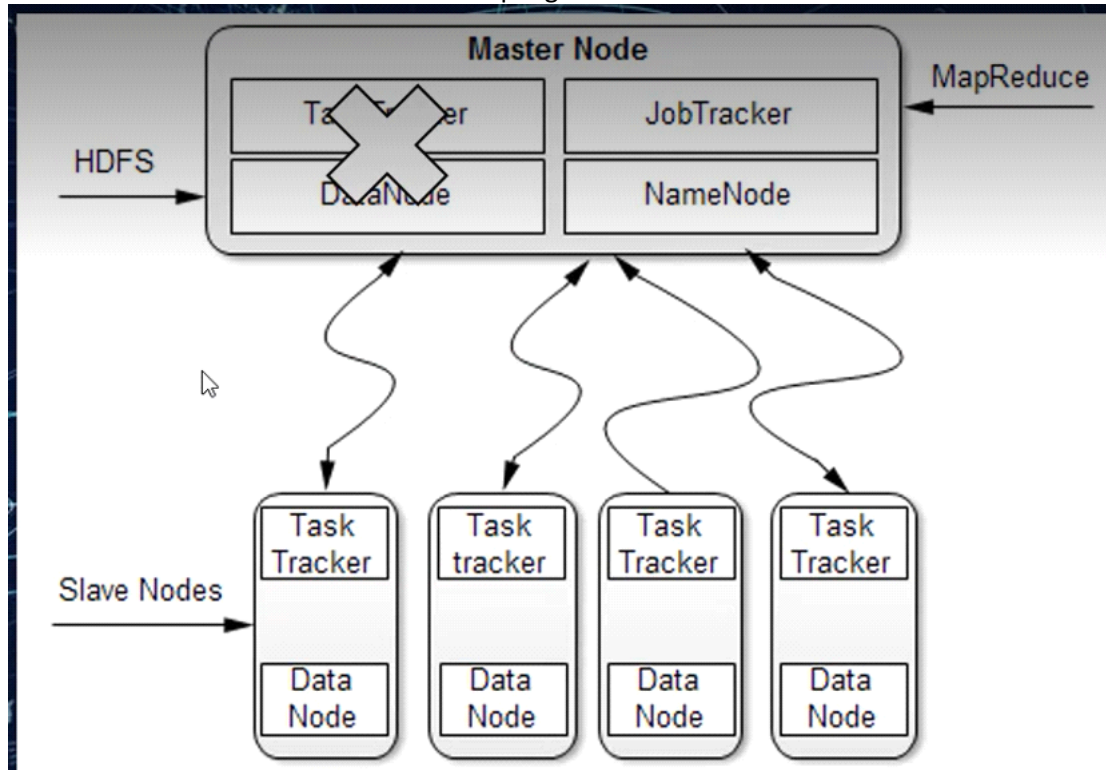
The Problem

- Single Central Storage
- Serial Processing – 1 I/P, processor, O/P

The Solution

- Distributed Storage
 - Parallel Processing
- **HADOOP**
(Multiple I/P, processor, O/P)

- We can use EMR of AWS for Hadoop Big Data. And in Azure data bricks in Microsoft Azure.



- Job tracker tracks the Slave nodes and the sub process that runs in the slave nodes.

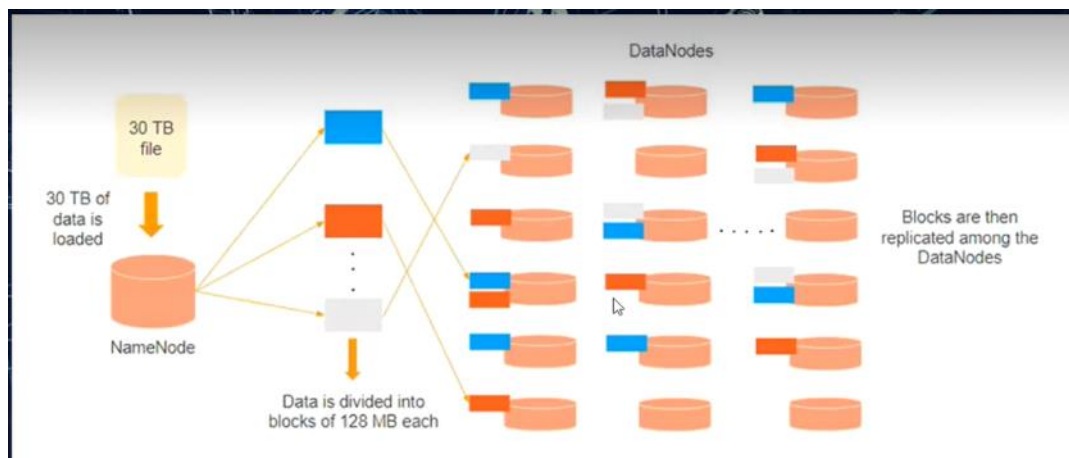
Master Node – Java process = Name node

Slave Node – Java process = Data node

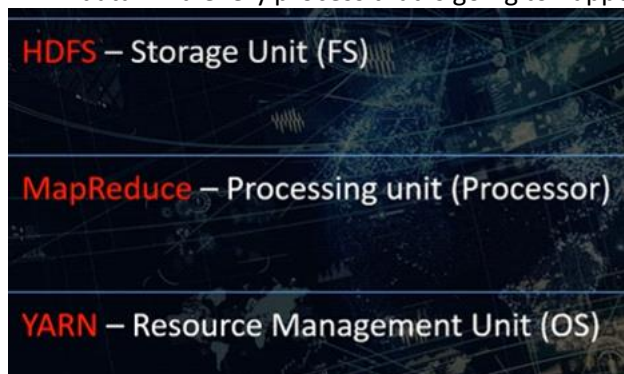
Name Node = Stores metadata, Receives heartbeat

Sec Name Node = Edit Log & FS Image

Data Node = Read/ write/ process/ replicate



- Hadoop internally replicate the data on several data nodes because to prevent the loss of data. And every process that is going to happen on data nodes will be located in log files.



Function of Map Reduce:



- First we divide into chunks and then map it and then reduce it and then sum it and merge it. And return that result to the master node.

ETL

Monday, March 28, 2022 7:37 PM

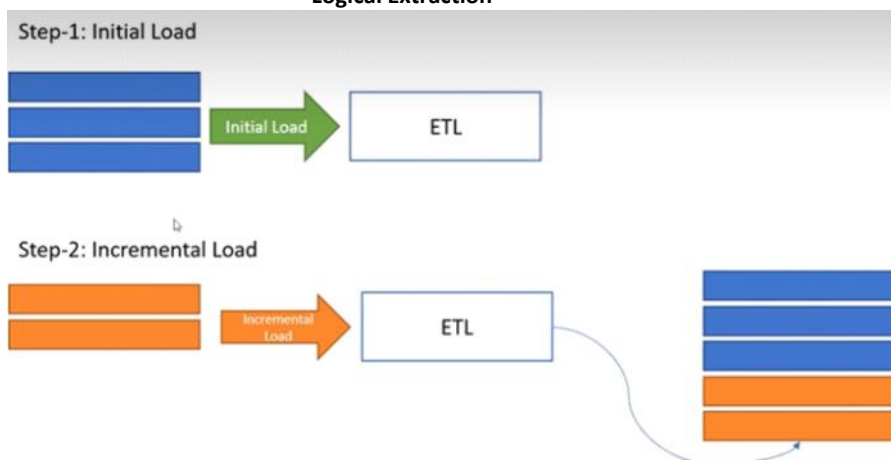


- As every application runs on the data , the ETL Process is necessary.
- For any company that requires data , we collect it from many sources and using that data we transform that data into the way that we want. And load that data to store on our databases.
- Whenever we have data , we have to do ETL process.
- For pipelining we use other tools other than ETL tools.
- Whenever we are considering data, we need to consider data source, data types.

Extraction

- We can find source data at any where on cloud or on any data source.
- In general every dataset supports row type format. But in general in those conditions , if we want to retrieve data from a particular column then we must scan all the rows and use only one column of that row data. It means we are loading 9 extra data in 10 to retrieve only 1 column data. So for that case , now a days we are introducing column format data.
- In those column format data, the column data is stored in a single row such that if we want to retrieve data of that column then we just need to import only one row.
- Data in Nielsen is stored in snowflake database.
- For storing column based data format snowflake is used.
- There are two types of extracting the data, physical extraction and logical extraction.

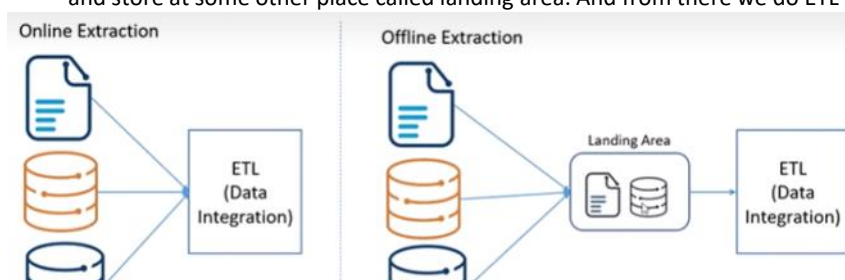
Logical Extraction



- In incremental load model , we don't load the previously designed data repeatedly. We just add the new data to already existing data.

Physical extraction

- **Online extraction:** we need to collect data from clients cloud directly through online. Due to this all the main security data and location data will be leaked to others.
- **Offline extraction:** To prevent such leak we use offline extraction, where we carries the data and store at some other place called landing area. And from there we do ETL process.





- In general , everyone prefers offline extraction. It depends on cost and sensitiveness of the data.
- We do extracting and loading tools by the following technologies.
- In Nielsen we connect to the snowflake database and then we retrieve the tables that we require to azure blob storage.
- In Nielsen we retrieve the database from snowflake and perform the transform operations and resend it to the snowflake database.
- **Tools for extraction:**



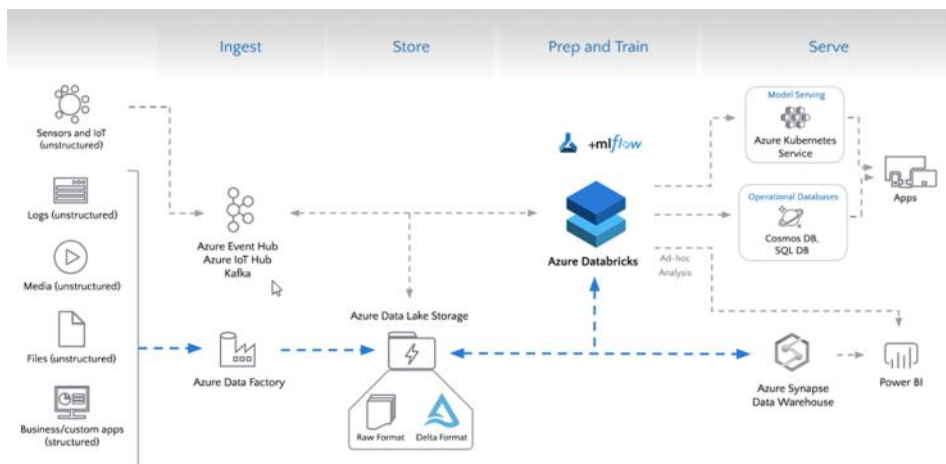
Transformation

- A simple change in the database or the files that contains the database even called as transformation.
- ETL tools in Nielsen:
 - For orchestration and extraction - airflow
 - For transformation - Databricks (In data bricks we use spark)
- **Tools for transformation:**

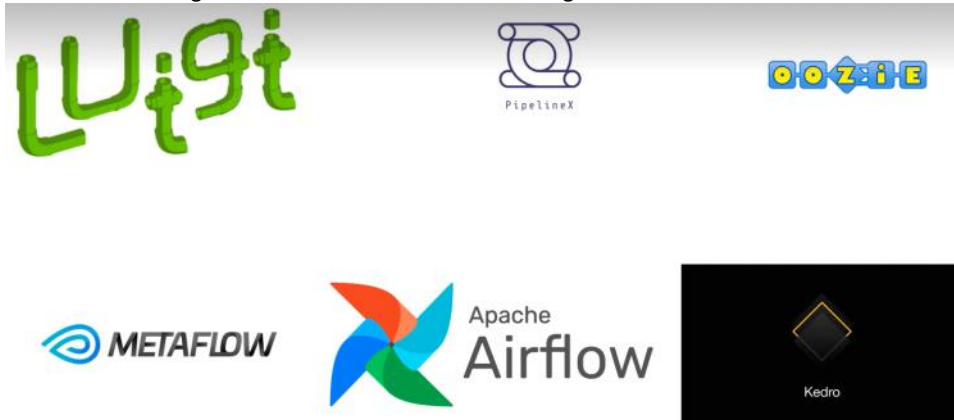


Orchestration

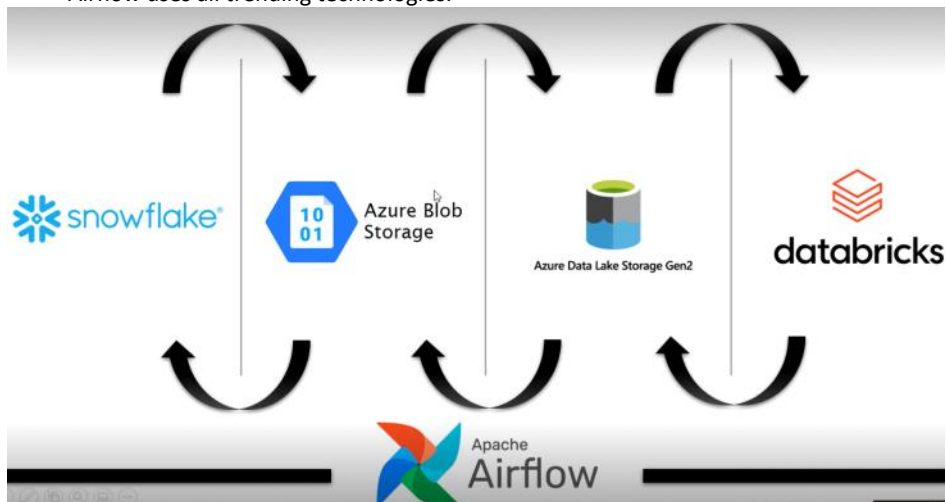
- For pipelining or scheduling or job chaining operations we use orchestration
- Using azure data factory we do extraction of the data.
- In Azure data lake storage , we store the data either in raw format or azure data format. And after storing we process the data using data bricks . And the new data is sent to the data ware house to store and the final report or analysis is made by PowerBI .



- For retrieving data from IOT devices we are having Azure Kafka.



- Airflow uses all trending technologies.



- Airflow monitors all the above operations.
- Snowflake works on cluster instances.

AIRFLOW

- Front end - Angular, react .
- Most of the people for front end use angular because of its community support.
- Mainly react concentrate on the mobile applications , it doesn't have more community support for web side of react JS.
- To prevent boiler plate code or to reduce the code all these new type of technologies are introducing.
- Python have many applications for any kind of the developer.
- Python 3 doesn't support previous python versions.
- NoSQL is trending these days.
- Shift is Datawarehouse for AWS
- Synapse is Datawarehouse for azure
- Bit query is Datawarehouse for GCP
- But snowflake can be used for any cloud platform as per our choice.
- For testing we use selenium.
- For DevOps we use docker and Kubernetes.

- We use AIRFLOW in Kubernetes.

APACHE AIRFLOW

"A platform created by community to programmatically author, schedule and monitor workflows."

- ⌚
It's a job scheduler that run your tasks by obligating their execution dependencies.
- 📄
Workflows are written as code in the form of Directed Acyclic Graphs (DAGs)
- 🔄
Since workflows are written in DAGs, so they become more dynamic, manageable, testable, and collaborative.
- 🔧
Developed in Python so capable to interface with any third party python API and can execute an endless variety of tasks irrespective of their language.

Basic Terminologies in Airflow

DAG

A DAG (Directed Acyclic Graph) is unidirectional, acyclic graph connecting the edges, where each node in the graph is a task, and edges define dependencies amongst tasks.

Operators

- An operator represents a single task in a workflow that helps to carry out your task.
- Operators determine what actually gets to be done when your DAG runs.

Task

Task is an operator when instantiated. It is something on which the worker works upon.

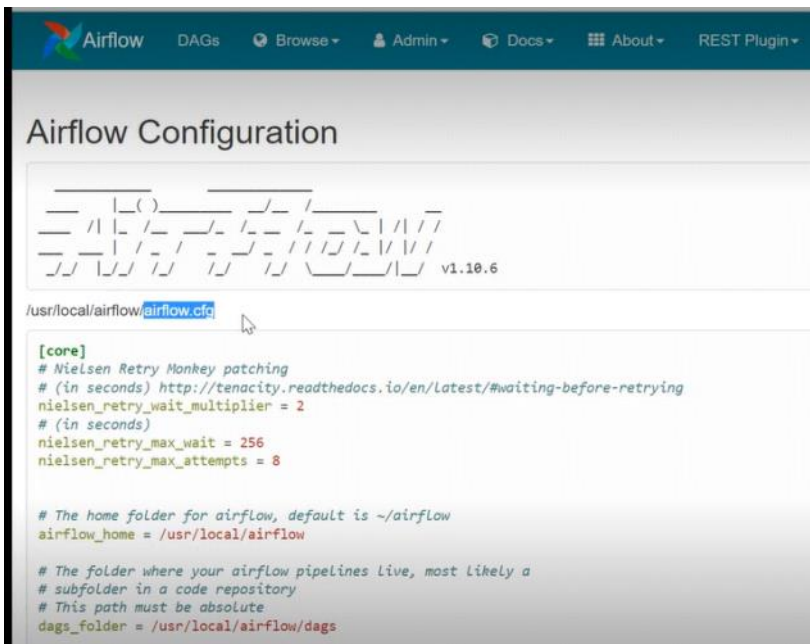
Workflow

Workflow is a sequence of tasks arranged in a control dependency. Workflow and DAG can be used interchangeably.

- For every operator that we want to execute, we have a operator. For python we have python operator , for data bricks we have data bricks operator. We need to import that operator and utilize that.



- We can give such links between operations and all these graphs , tables were created by using user code under code section.
- We could not create the graphs in drag and drop manner, it must be designed using code.



- If we want to configure airflow on our local system we need to modify airflow.cfg , where we can modify how many times we can try when our current operation failed and etc.
- If we use Kubernetes executor use parameter value of executor as Kubernetes.
- Even we can assign the configuration file to send the errors to the mail id using smtp server.

```
[email]
email_backend = airflow.utils.email.send_email_smtp

[smtp]
# If you want airflow to send emails on retries, failure, and you want to use
# the airflow.utils.email.send_email_smtp function, you have to configure an
# smtp server here
smtp_host = smarthost.enterprisenet.org
smtp_starttls = false
smtp_ssl = false
smtp_user =
smtp_port = 25
smtp_password =
smtp_mail_from = airflowadmin@nielsen.com
```

- We can change kubernetes details

```
[kubernetes]
# The repository, tag and imagePullPolicy of the Kubernetes Image for the Worker to Run
worker_container_repository = buycscregprod.azurecr.io/buy-ndx/ndx-airflow
worker_container_tag = 1.5.6-105
worker_container_image_pull_policy = IfNotPresent

# If True (default), worker pods will be deleted upon termination
delete_worker_pods = True

# The Kubernetes namespace where airflow workers should be created. Defaults to `default`
namespace = airflow

# The name of the Kubernetes ConfigMap Containing the Airflow Configuration (this file)
airflow_configmap = airflow-configmap

# For either git sync or volume mounted DAGs, the worker will Look in this subpath for DAGs
#dags_volume_subpath = /usr/local/airflow/dags

# For DAGs mounted via a volume claim (mutually exclusive with volume claim)
dags_volume_claim = airflow-nfs-server-dags

# For volume mounted Logs, the worker will Look in this subpath for Logs
#logs_volume_subpath = /usr/local/airflow/logs

# A shared volume claim for the Logs
logs_volume_claim = airflow-nfs-server-logs
```

- We can use configuration file to link our git repository to airflow.

```
# For volume mounted Logs, the worker will Look in this subpath for Logs
#logs_volume_subpath = /usr/local/airflow/logs

# A shared volume claim for the Logs
logs_volume_claim = airflow-nfs-server-logs

# Git credentials and repository for DAGs mounted via Git (mutually exclusive with volume claim)
git_repo = ssh://git@adlm.nielseniq.com:7999/drd/kBs-airflow-dag-nprod.git
git_branch = master
git_user =
git_password =
git_subpath =
```

```
# For volume mounted logs, the worker will look in this subpath for logs
#logs_volume_subpath = /usr/local/airflow/logs

# A shared volume claim for the logs
logs_volume_claim = airflow-nfs-server-logs

# Git credentials and repository for DAGs mounted via Git (mutually exclusive with volume claim)
git_repo = ssh://git@adlm.nielseniq.com:7999/drd/k8s-airflow-dag-nprod.git
git_branch = master
git_user =
git_password =
git_subpath =

# For cloning DAGs from git repositories into volumes: https://github.com/kubernetes/git-sync
git_sync_container_repository = gcr.io/google-containers/git-sync-amd64
git_sync_container_tag = v2.0.5
git_sync_init_container_name = git-sync-clone
```

- In variable section of airflow we can store the key with a value , such as passwords or some important information which we don't want to hard code it repeatedly we use this variables. It is similar to a normal variable but it will be in key value format.

No file chosen

List Variable

Search

«

<

0

1

2

3

4

5

6

>

»

Page size

+

Actions

←

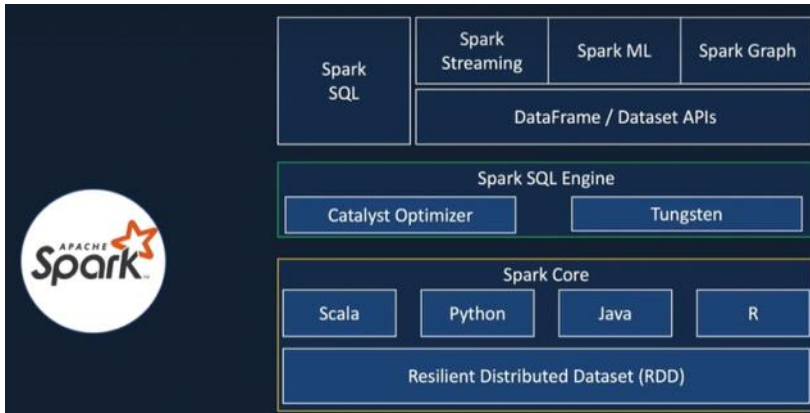
Record Count: 309

		Key	Val	Is Encrypted
<input type="checkbox"/>	<input checked="" type="checkbox"/>	adhoc_client_secret	*****	True
<input type="checkbox"/>	<input checked="" type="checkbox"/>	adhoc_token_password	*****	True
<input type="checkbox"/>	<input checked="" type="checkbox"/>	airflow_db_cleanup__max_db_entry_age_in_days	90	True

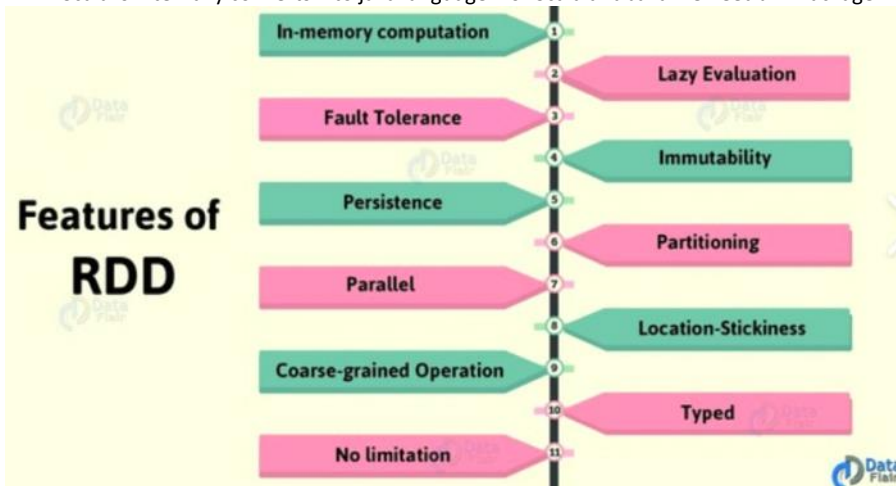
Data Bricks

Tuesday, March 29, 2022 10:44 AM

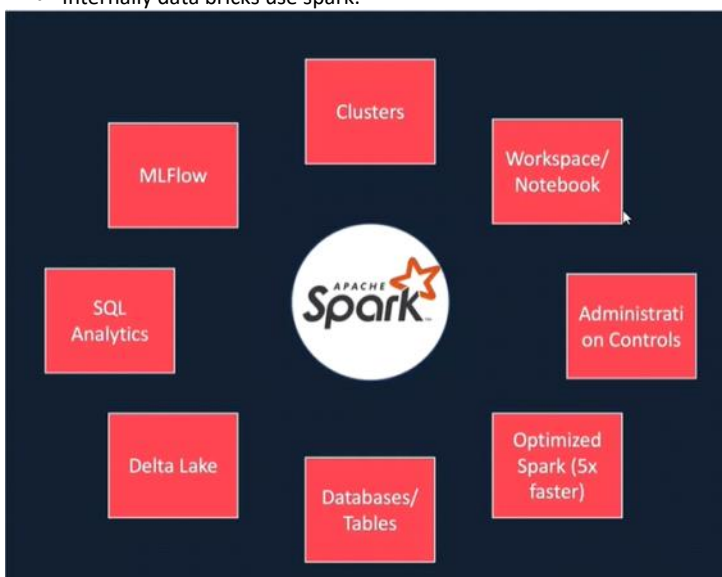
- Spark using data bricks is preferable.
- Once if we don't get a heartbeat from slave node then master node waits for 3 seconds and then master node recovers the data into processing. Hadoop and spark are fault tolerant due to this method.



- Mostly in spark we use Scala language because spark is designed by Scala language. But if we want to use complex machine learning algorithms we can use python language. There is no restriction of language.
- Scala is internally converts into java language. For Scala and Java we need JRE Package.

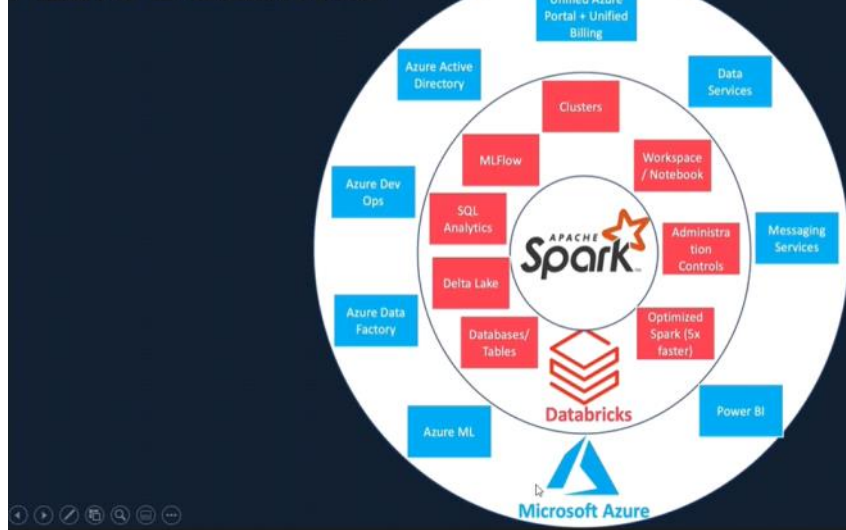


- Internally data bricks use spark.



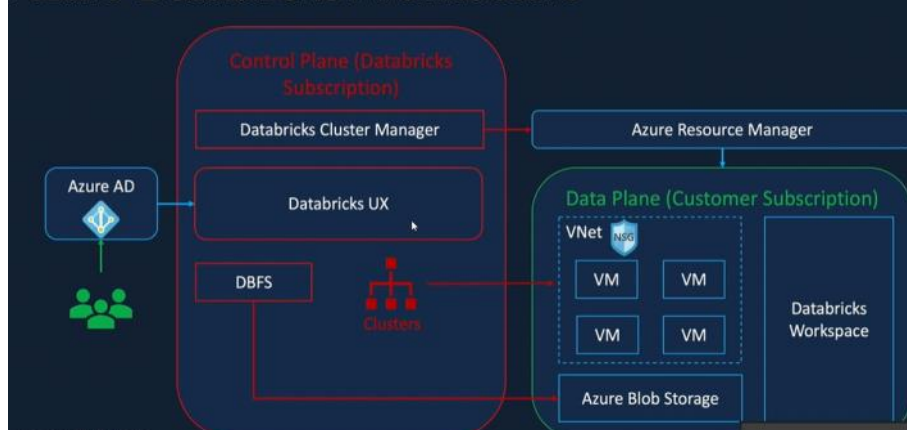
- Internally if we use spark in data bricks operations will be optimized 5 times faster.

Azure Databricks

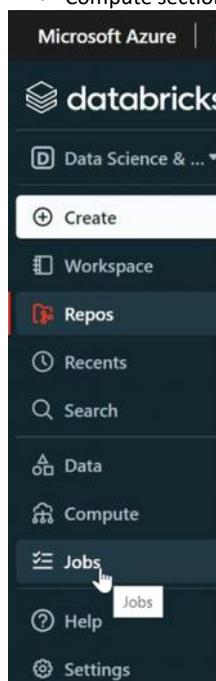


- Databricks is independent service but we can use data bricks internally of Microsoft Azure.

Azure Databricks Architecture



- DBFS is a small database in the architecture. And to implement it again we use azure blob storage.
- Databricks is integrated with azure. It is a platform as a service.
- Compute section of azure data bricks is used as cluster manager



Microsoft Azure | Databricks Portal anton.milan@nielseniq.com

Compute

All-purpose clusters Job clusters Pools

Create Cluster

All Created by me Accessible by me Filter...

Name	State	Nodes	Runtime	Driver	Worker	Creator	Actions
Swagger_Test	Terminated	-	5.5 LTS (includes Apache Spark 2.4.3, Scala...	Standard_...	Standard_...	csdbsvacant...	0

1 - 1 of 1 20 / Page Go to 1

- We will schedule these jobs on airflow , when job is called cluster is triggered here and start the cluster.
- On azure data blob storage we could not do more processing as compared to azure data lake storage.
- Sc command is used to access RDD (Spark context)
- Spark command is used to access Spark SQL. (Spark session)

To Create an RDD (Resilient distributed dataset):

```
Cmd 2
1 val rdd = sc.textFile("dbfs:/FileStore/tables/wc.txt")
rdd: org.apache.spark.rdd.RDD[String] = dbfs:/FileStore/tables/wc.txt MapPartitionsRDD[1] at
Command took 0.53 seconds -- by anton.milan@nielseniq.com at 3/29/2022, 11:58:25 AM on Swagger_Test
```

Print first line through RDD

```
1 //val rdd = sc.textFile("dbfs:/FileStore/tables/wc.txt")
2 rdd.first()

(1) Spark Jobs
res7: String = car bus lorry
Command took 2.94 seconds -- by anton.milan@nielseniq.com at 3/29/2022, 11:59:03 AM on Swagger_Test
```

Now we need to explode the sentence, for that we use map function and here take function is used to print first 10 records of that returned map list. In current example , we are having only four lines. But we just tried to print first 10 lines.

Flat map maps and returns a single array

```
Cmd 2
1 //val rdd = sc.textFile("dbfs:/FileStore/tables/wc.txt")
2 rdd.flatMap(x => x.split(" ")).take(15)

(2) Spark Jobs
res12: Array[String] = Array(car, bus, lorry, auto, car, bus, bike, lorry, auto, car, lorry, car)
Command took 0.34 seconds -- by anton.milan@nielseniq.com at 3/29/2022, 12:02:36 PM on Swagger_Test
```

Map function after flat map operation

```
Cmd 2
1 //val rdd = sc.textFile("dbfs:/FileStore/tables/wc.txt")
2 rdd.flatMap(x => x.split(" ")).map(x => (x,1)).take(15)

(2) Spark Jobs
res13: Array[(String, Int)] = Array((car,1), (bus,1), (lorry,1), (auto,1), (car,1), (bus,1), (ar,1))
Command took 0.48 seconds -- by anton.milan@nielseniq.com at 3/29/2022, 12:04:07 PM on Swagger_Test
```

Finally we do word count using RDD using FlatMap map and reducebykey and take operations on the four liner text file.

```
Cmd 2
1 //val rdd = sc.textFile("dbfs:/FileStore/tables/wc.txt")
2 rdd.flatMap(x => x.split(" ")).map(x => (x,1)).reduceByKey((x,y) => x+y).collect()

(1) Spark Jobs
res17: Array[(String, Int)] = Array((lorry,3), (car,4), (bus,2), (auto,2), (bike,1))
Command took 0.52 seconds -- by anton.milan@nielseniq.com at 3/29/2022, 12:11:18 PM on Swagger_Test
```

Cmd 2

```

1 //val rdd = sc.textFile("dbfs:/FileStore/tables/wc.txt")
2 rdd.flatMap(x => x.split(" ")).map(x => (x,1)).reduceByKey((x,y) => x+y).collect().foreach(println)

```

► (1) Spark Jobs

```

(lorry,3)
(car,4)
(bus,2)
(auto,2)
(bike,1)

```

Command took 0.52 seconds -- by anton.milan@nielseniq.com at 3/29/2022, 12:11:53 PM on Swagger_Test

Data Frame

Cmd 3

```

1 //l df= spark.read.textFile("dbfs:/FileStore/tables/wc.txt").toDF("lines")
2 df.select($"lines").show()
3

```

► (1) Spark Jobs

```

+-----+
|      lines|
+-----+
| car bus lorry|
|  auto car bus|
|bike lorry auto|
|  car lorry car|
+-----+

```

Cmd 3

```

1 //l df= spark.read.textFile("dbfs:/FileStore/tables/wc.txt").toDF("lines")
2 import org.apache.spark.sql.functions._
3
4 df.select(split(df("lines"), " ")).show()
5

```

► (1) Spark Jobs

```

+-----+
| split(lines, )|
+-----+
| [car, bus, lorry]|
| [auto, car, bus]|
| [bike, lorry, auto]|
| [car, lorry, car]|
+-----+

```

Cmd 3

```

1 //l df= spark.read.textFile("dbfs:/FileStore/tables/wc.txt").toDF("lines")
2 import org.apache.spark.sql.functions._
3
4 df.select(split(df("lines"), " ").alias("words")).show()
5

```

► (1) Spark Jobs

```

+-----+
|      words|
+-----+
| [car, bus, lorry]|
| [auto, car, bus]|
| [bike, lorry, auto]|
| [car, lorry, car]|
+-----+

```



```

Cmd 3
1 //l df= spark.read.textFile("dbfs:/FileStore/tables/wc.txt").toDF("lines")
2 import org.apache.spark.sql.functions._
3
4 df.select(explode(split(df("lines"), " ")).alias("words")).groupBy("words").count().show()
5
  (5) Spark Jobs
+-----+-----+
| words | count |
+-----+-----+
|  auto |     2 |
|   bus |     2 |
| lorry |     3 |
|   car |     4 |
|  bike |     1 |
+-----+-----+

```

Now a days everyone are using spark SQL instead of RDD and Spark Data frame.
We can even execute SQL operations on the text file or on the data frames using following manner

```

Cmd 4
1 //spark.read.textFile("dbfs:/FileStore/tables/wc.txt").toDF("lines").createTempView("wc")
2 spark.sql("select * from wc").show()

  (1) Spark Jobs
+-----+
|      lines |
+-----+
| car bus lorry |
| auto car bus |
| bike lorry auto |
+-----+

```

We can even use SQL operations

```

Cmd 4
1 //spark.read.textFile("dbfs:/FileStore/tables/wc.txt").toDF("lines").createTempView("wc")
2 spark.sql("select words,count(*) from(select explode(split(lines, ' ')) words from wc) innerQ group by words").show()

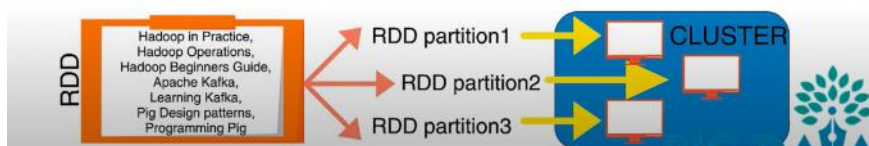
  (5) Spark Jobs
+-----+-----+
| words | count(1) |
+-----+-----+
|  auto |         2 |
|   bus |         2 |
| lorry |         3 |
|   car |         4 |
|  bike |         1 |
+-----+-----+

```

Here real tables will not be created, it mean we even no need to store currently on the Scala . We just create a meta data and when we perform any action we allocation the memory.
In above example, table will not be created , it is just a view of the table.

WHAT IS RDD ?

- ✓ RDD is the spark's core abstraction which is resilient distributed dataset
- ✓ It is the immutable distributed collection of objects
- ✓ Internally spark distributes the data in RDD, to different nodes across the cluster to achieve parallelization



RDD CAN BE CREATED IN 2 WAYS

By loading an external dataset

for example, loading an external dataset books.txt can be done as below

```
val booksRDD = sc.textFile("/path/to/books.txt")
```

By distributing collection of objects

for example, let's create a list collection and pass it to parallelize method of spark context

```
val colorsRDD = sc.parallelize(List("red", "blue"))
```

Load RDD in two different ways

```
scala> val booksRDD = sc.textFile("/home/hduser/mat/books.txt");
booksRDD: org.apache.spark.rdd.RDD[String] = MapPartitionsRDD[18] at textFile at <console>:15
scala>
scala> booksRDD.collect();
res10: Array[String] = Array(Hadoop in Practice, Hadoop Operations, Hadoop Beginners Guide, Apache Kafka,
Learning Kafka, Pig Design Patterns, Programming Pig)
scala>
RDD creation by distributing collection
scala> val colorsRDD = sc.parallelize(List("Red", "Blue"));
colorsRDD: org.apache.spark.rdd.RDD[String] = ParallelCollectionRDD[19] at parallelize at <console>:15
scala>
scala> colorsRDD.collect();
res11: Array[String] = Array(Red, Blue)
scala>
```