Microsoft

GETTING STARTED WITH HADOOP AND AZURE HDINSIGHT

Niti Gupta

AZURE HDINSIGHT

Niti Gupta

@nitigupta_16

@https://www.linkedin.com/in/niti-gupta-33200086

OUTLINE

- What is big data?
- Why Hadoop?
- What is HDInsight?
- Why HDInsight?
- Scenarios where you can use HDInsight.
- How can you get started?

AZURE HDINSIGHT

FUN FACTS

Every human created about 1.7 mb of data per second in 2020

(Tech Jury)

+

- WhatsApp users exchange up to 65 billion messages daily

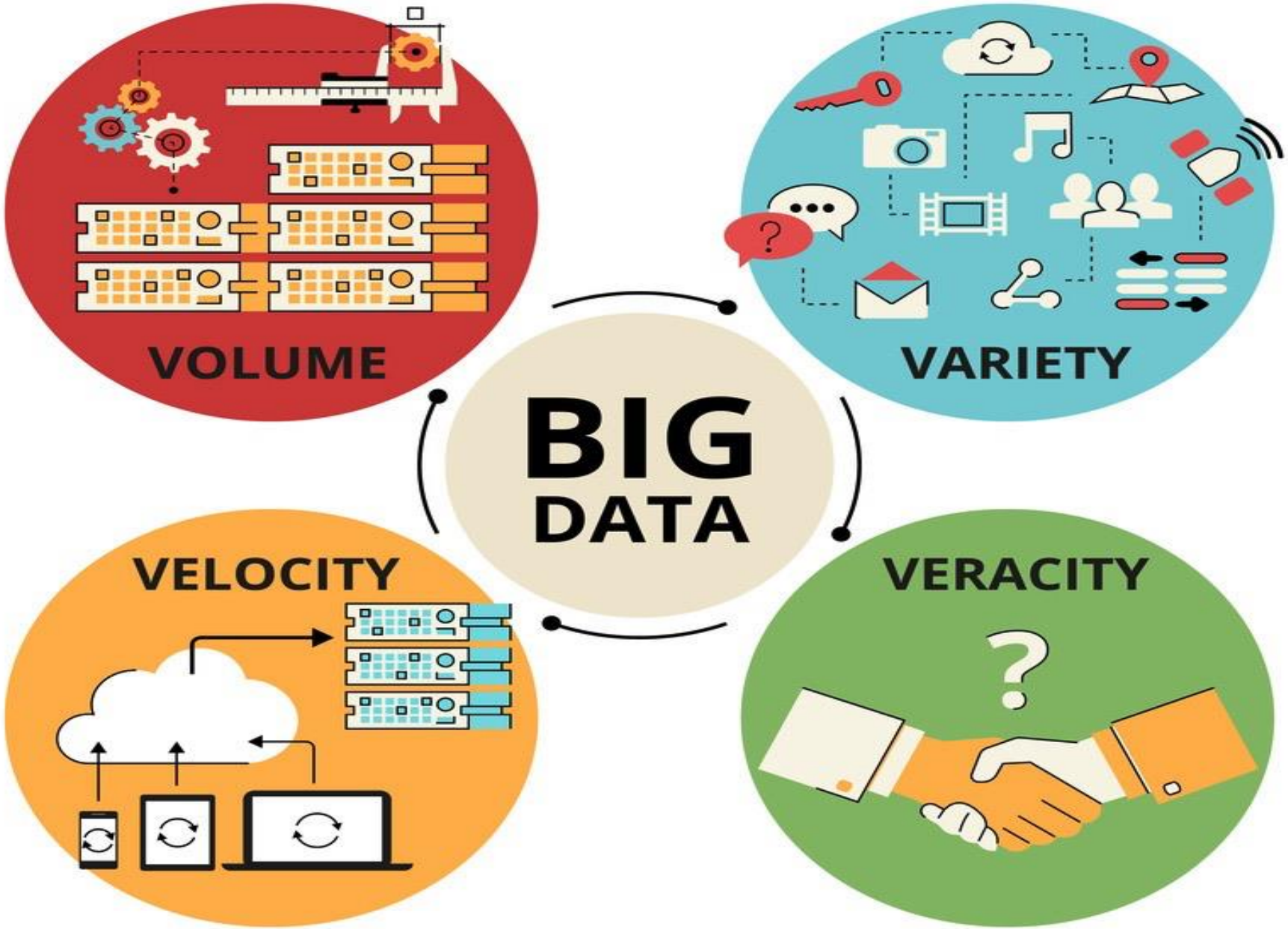
(Source: Connectiva Systems)

Every 2 days we create as much information as we did from the beginning of time until 2003

(CloudNine)

If you burned all of the data created in just one day onto DVDs, you could stack them on top of each other and reach the moon – twice

(CloudNine)


VOLUME
VARIETY
BIG
DATA
VELOCITY
VERACITY
?

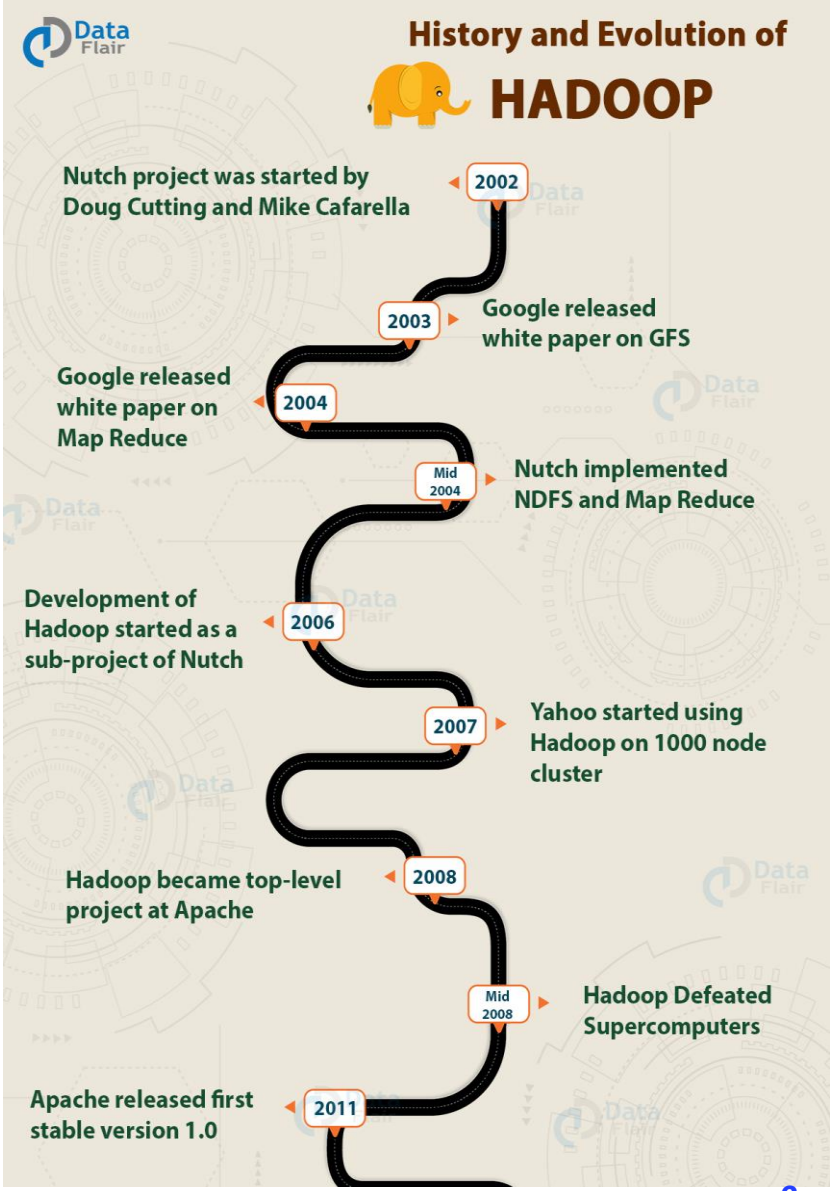
What is Hadoop?

Google's Solution

Google solved this problem using an algorithm called MapReduce. This algorithm divides the task into small parts and assigns them to many computers, and collects the results from them which when integrated, form the result dataset.

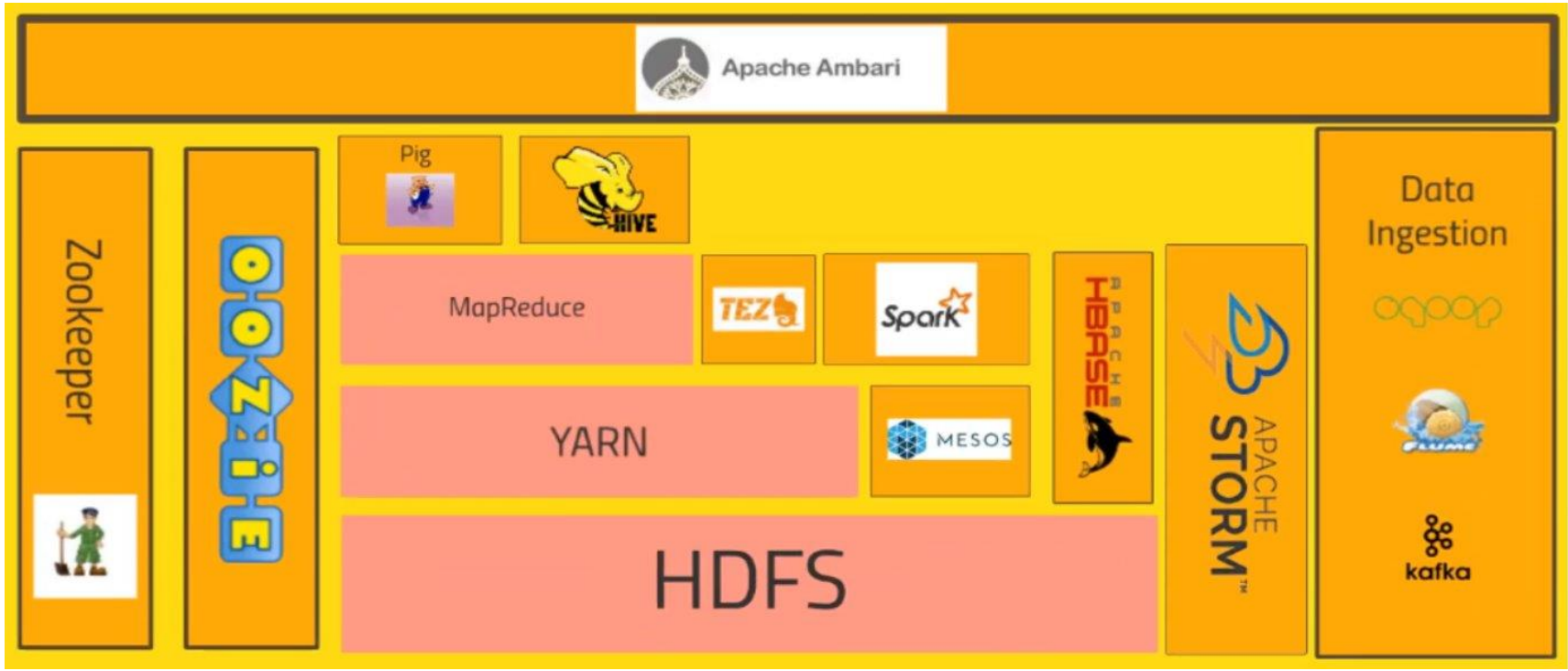
Using the solution provided by Google, Doug Cutting and his team developed an Open Source Project called **HADOOP**.

Hadoop runs applications using the MapReduce algorithm, where the data is processed in parallel with others.

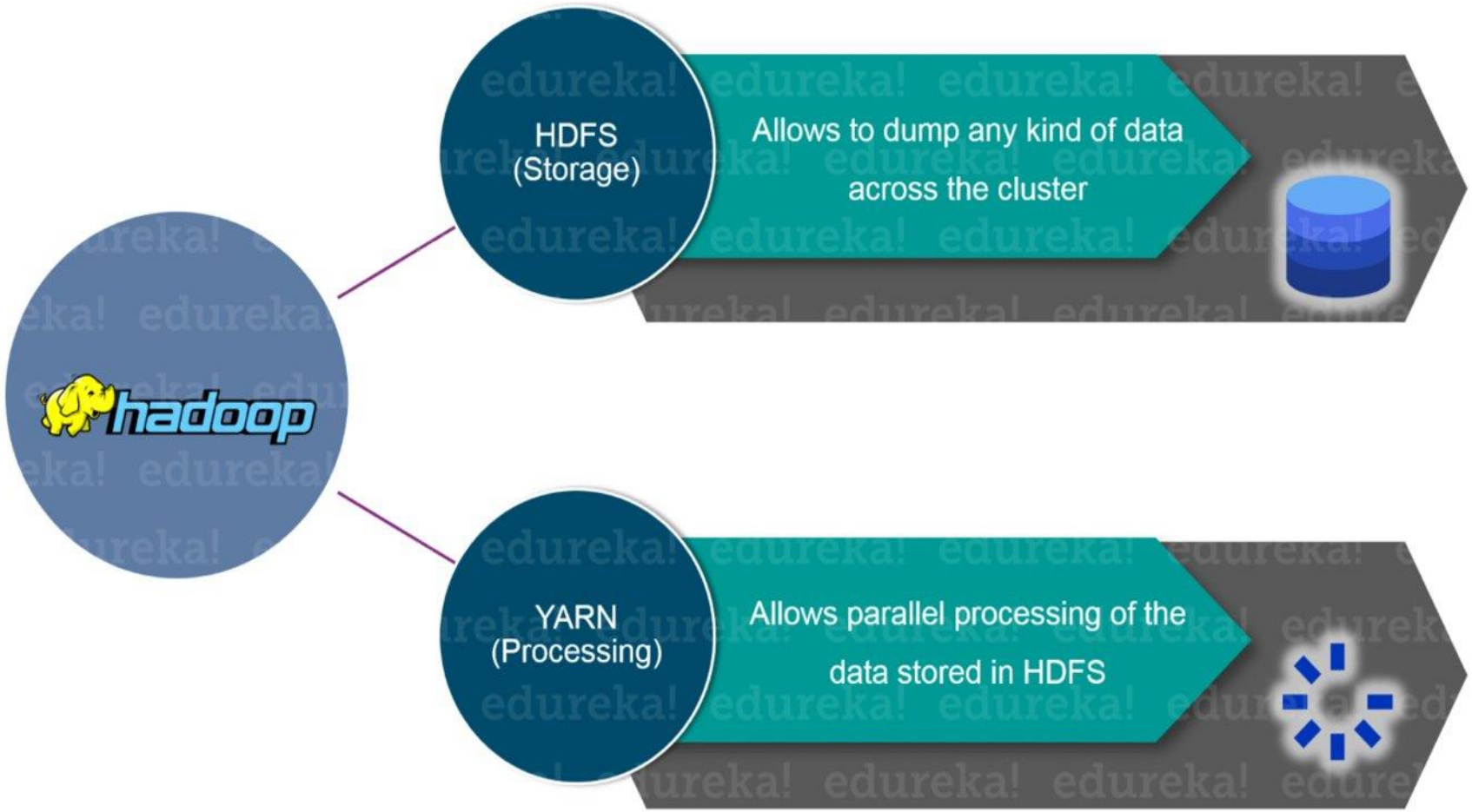


Discover Hadoop

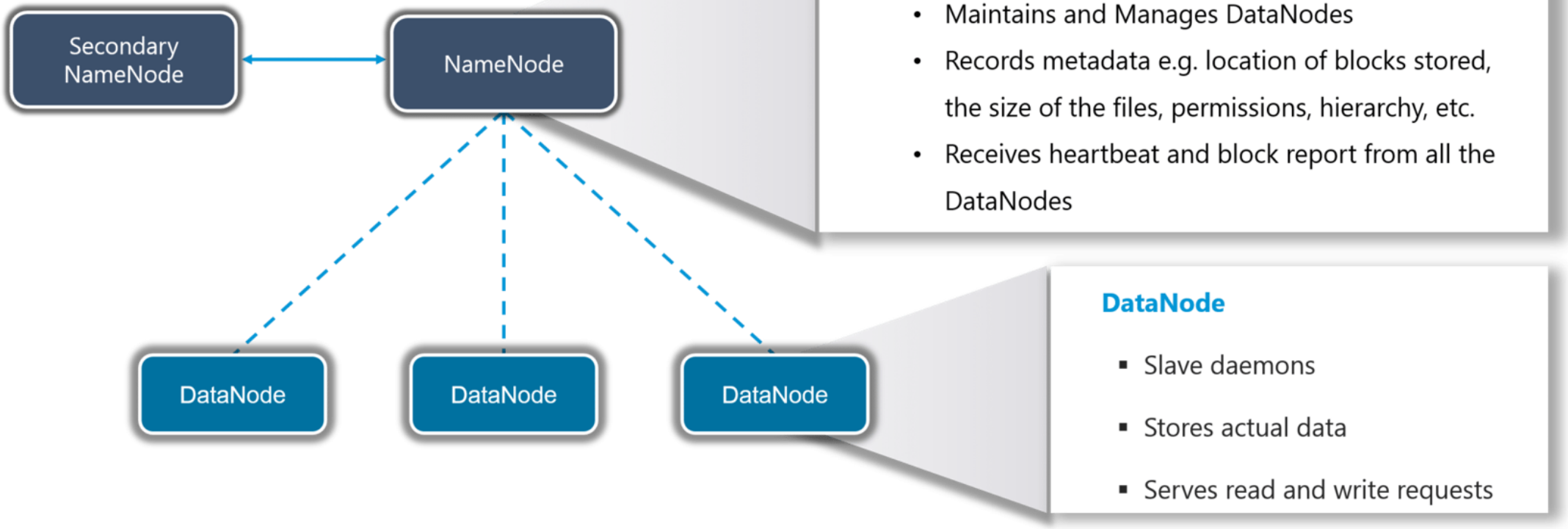
Hadoop is an open-source software framework for storing data and running applications on clusters of commodity hardware. It provides **massive storage** for any kind of data, **enormous processing power** and the ability to handle virtually **limitless concurrent tasks** or jobs.



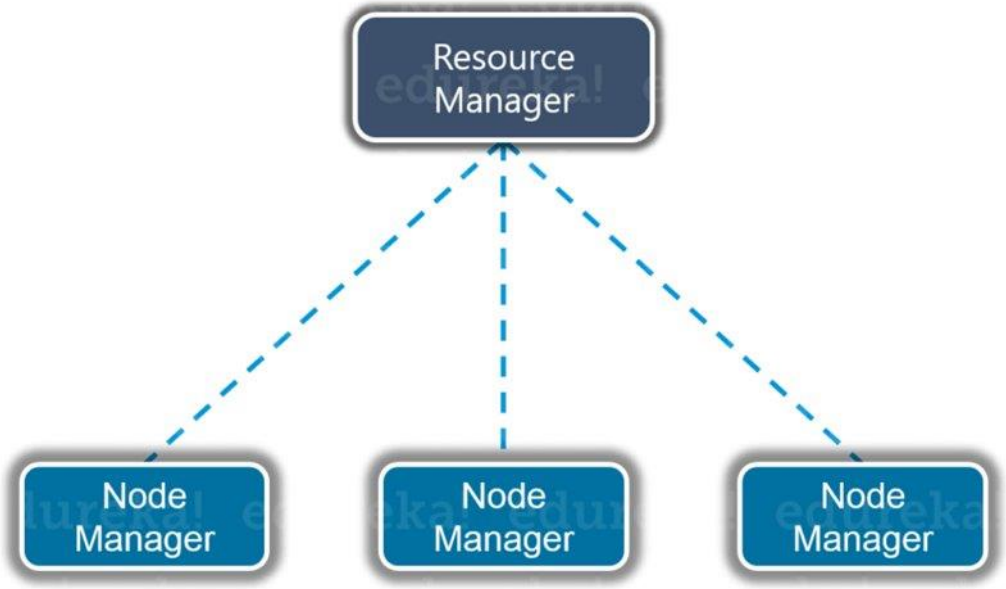
Hadoop Core Components



HDFS


Secondary NameNode
NameNode
DataNode
DataNode
DataNode
NameNode
Master daemon
Maintains and Manages DataNodes
Records metadata e.g. location of blocks stored, the size of the files, permissions, hierarchy, etc.
Receives heartbeat and block report from all the DataNodes
DataNode
Slave daemons
Stores actual data
Serves read and write requests

YARN



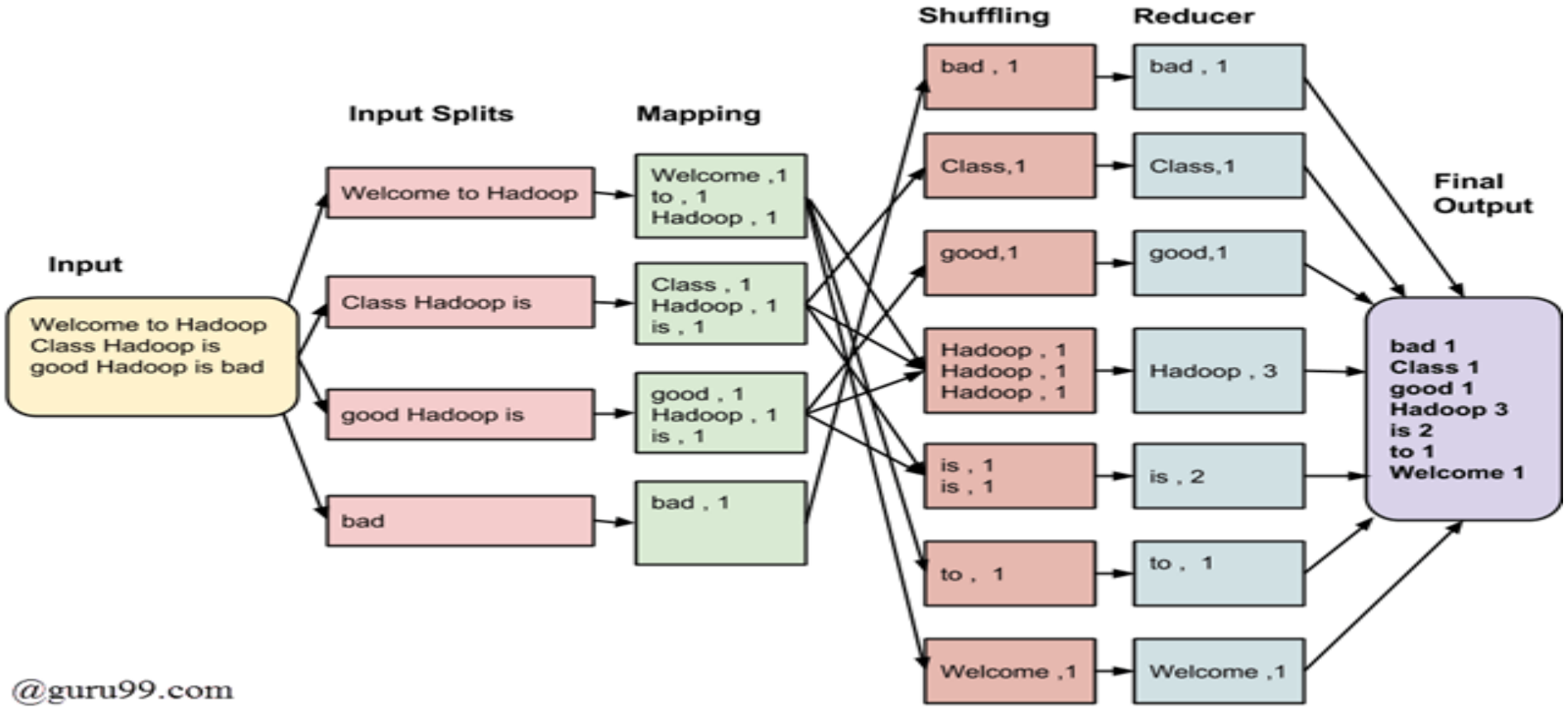
ResourceManager

- Receives the processing requests
- Passes the parts of requests to corresponding NodeManagers

NodeManagers

- Installed on every DataNode
- Responsible for execution of task on every single DataNode

MAP REDUCE


Input
Welcome to Hadoop
Class Hadoop is
good Hadoop is bad
Input Splits
Welcome to Hadoop
Class Hadoop is
good Hadoop is
bad
Mapping
Welcome ,1
to , 1
Hadoop , 1
Class , 1
Hadoop , 1
is , 1
good , 1
Hadoop , 1
is , 1
bad , 1
Shuffling
bad , 1
Class,1
good,1
Hadoop , 1
Hadoop , 1
Hadoop , 1
is , 1
is , 1
to , 1
Welcome ,1
Reducer
bad , 1
Class,1
good,1
Hadoop , 3
is , 2
to , 1
Welcome ,1
Final
Output
bad 1
Class 1
good 1
Hadoop 3
is 2
to 1
Welcome 1
@guru99.com


HDINSIGHT A WHAT?

Azure HDInsight

- Managed, full-spectrum, open-source analytics service in the cloud.
- Cloud distribution of Hadoop components.
- Core Components of Hadoop like HDFS, YARN, MapReduce are used to analyze batch data in Hadoop cluster of HDInsight.
- Integration with open-source frameworks such as Hadoop, Spark, Hive, LLAP, Kafka, Storm, R, and more supporting multiple programming languages.

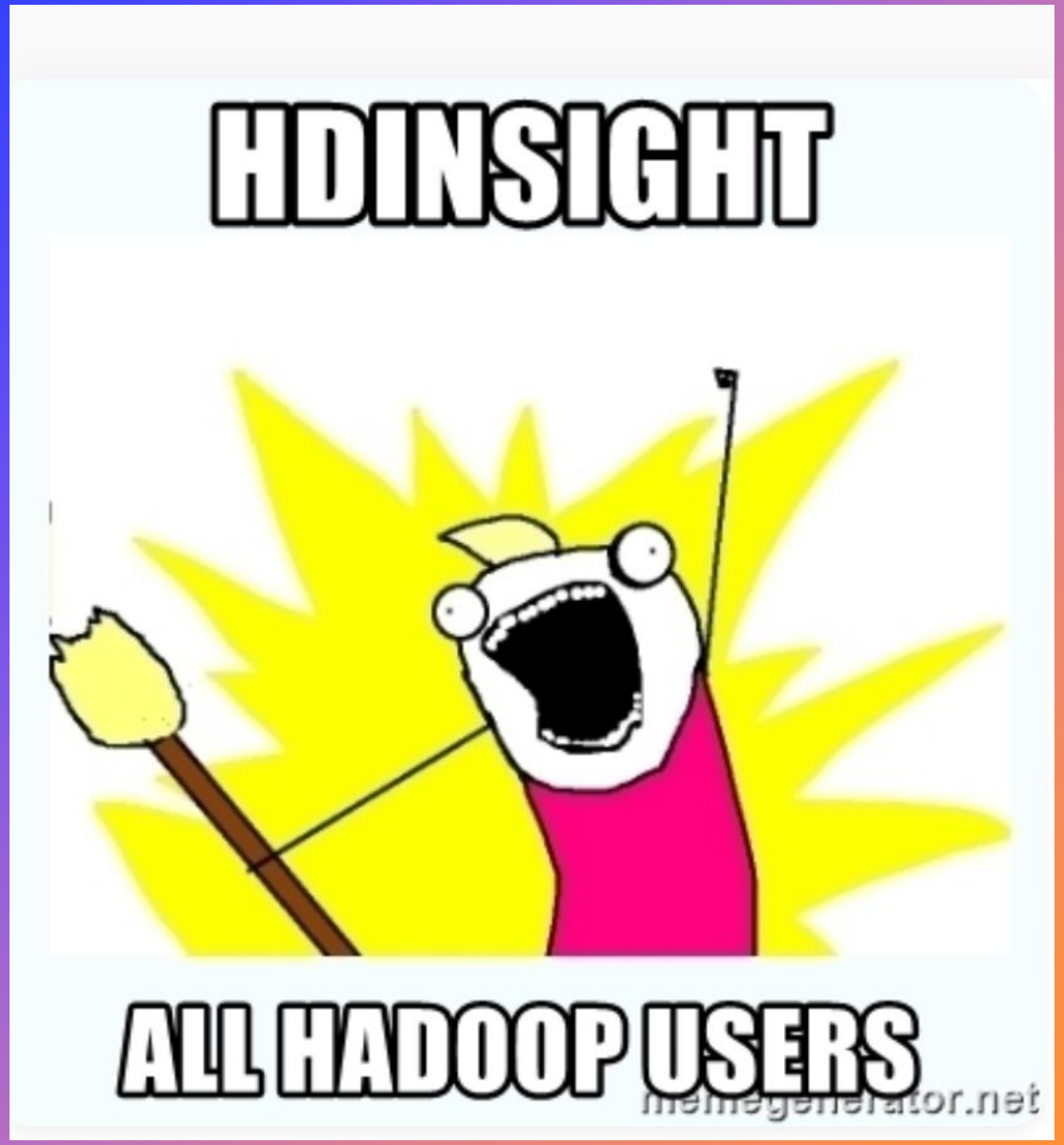
FEATURES OF HDINSIGHT

Reduced Costs

Secure & Compliant

Highly scalable

Optimized Components



HDInsight Advantages Over Hadoop

- Low Cost
- Automated Cluster Creation
- Monitoring
- Managed Hardware and Configuration
- Global Availability
- Security and Compliance
- Integration with other Azure Services like cosmos Db, blob storage, data factory
- Higher productive by using rich tools with Hadoop, Spark and Machine Learning

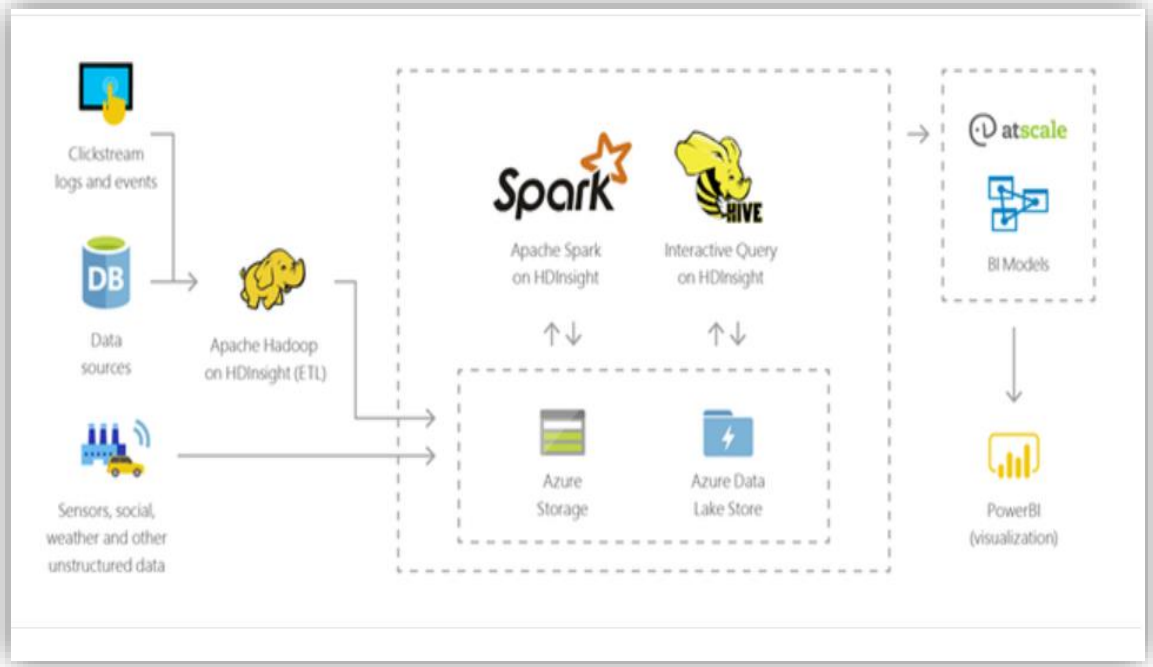
AZURE HDINSIGHT

Cluster Type	Functionality
Hadoop	Batch Query
HBase	Processing Schema free Data
Kafka	Streaming Platform
ML Services	Predictive Modelling
Spark	In-memory processing
Storm	Real-time event processing
Interactive Query	In-memory caching

Scenarios for using HDInsight

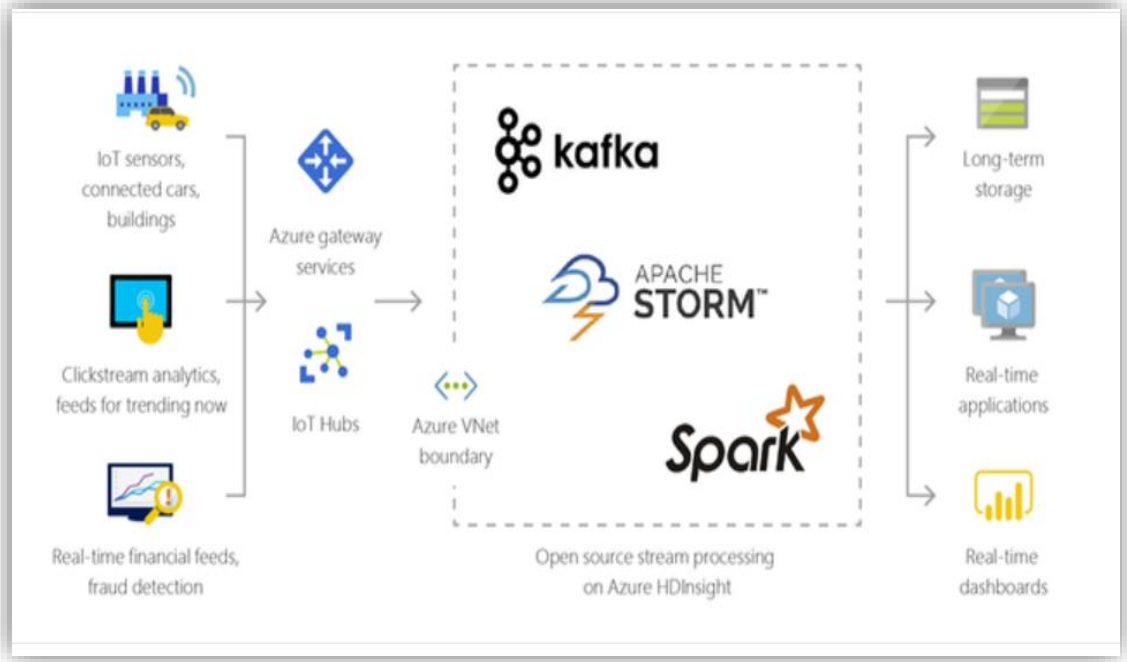
Batch Processing and data warehousing

- ETL at petabytes scale
- Interactive queries and models using BI capabilities



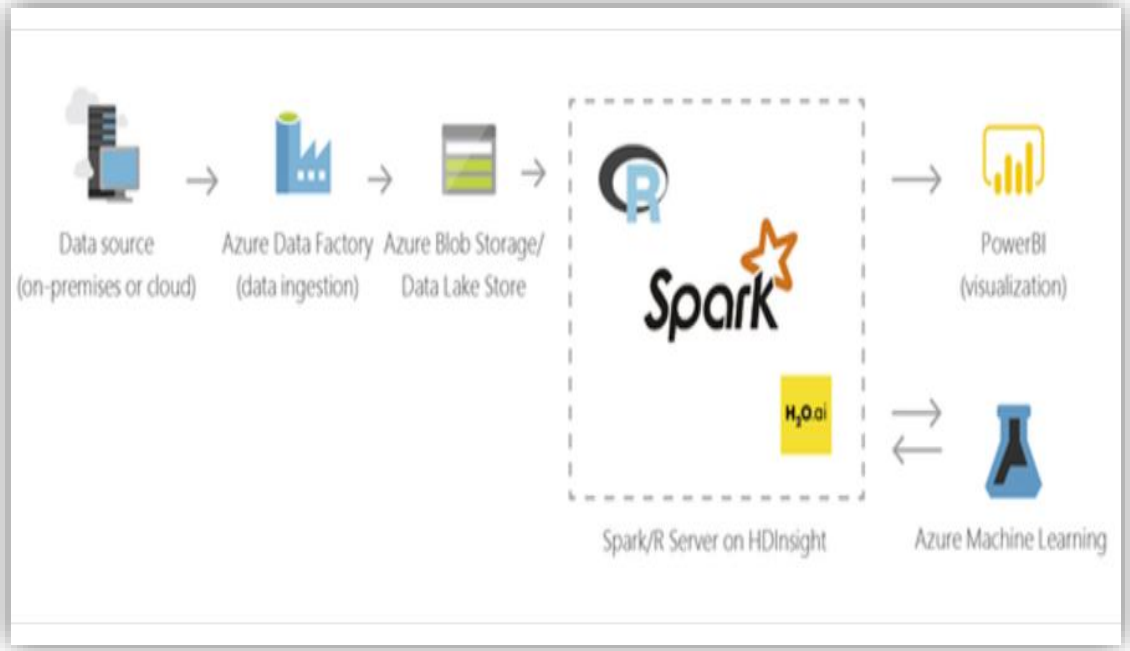
IoT

- Process streaming data in real-time received from different devices



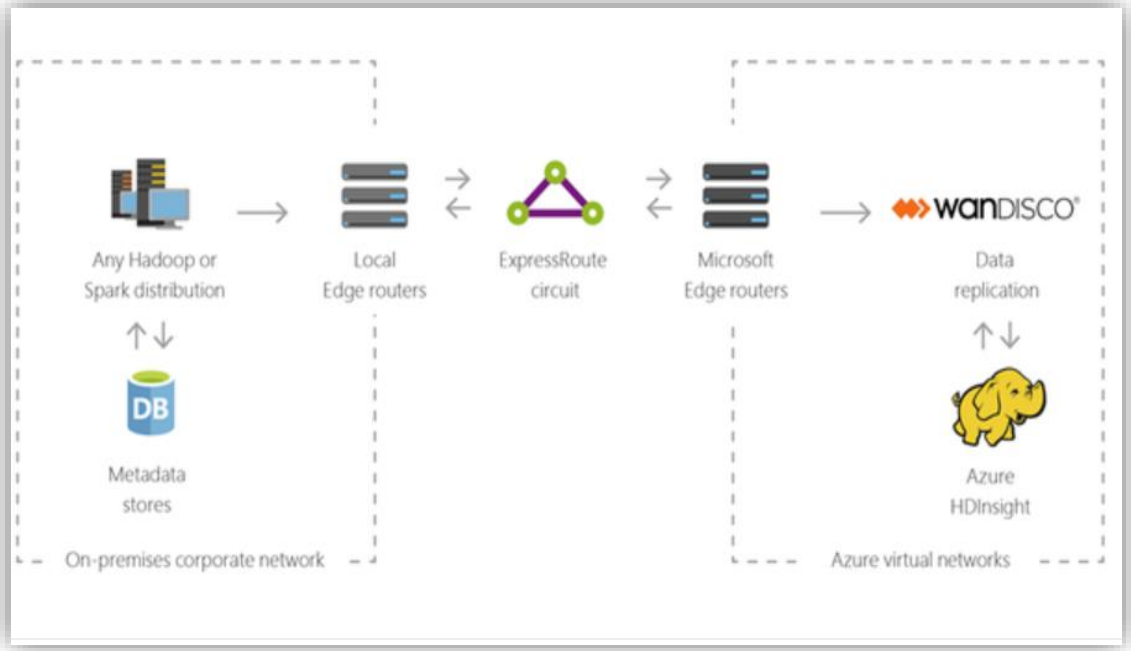
Data Science

- Extract critical insights from data
- Use Azure Machine learning to predict future trends



Hybrid

- Extend on-premise big data infrastructure to Azure to leverage cloud capabilities




wow
hdinsight
such idea
so easy
production

CLUSTER CREATION DEMO

Home > Create a resource > Marketplace >

Azure HDInsight

Microsoft

Azure HDInsight

Add to Favorites

Microsoft

Create

Create

Overview Plans Usage Information + Support Reviews

HDInsight is a Big Data service from Microsoft that brings 100% Apache Hadoop and other popular Big Data solutions to the cloud. A modern, cloud-based data platform that manages data of any type. Whether your data is structured or unstructured, and of any size, HDInsight makes it possible for you to gain the full value of Big Data.

With HDInsight, you can seamlessly process data of all types through Microsoft's modern data platform. Our platform provides simplicity, ease of management, and an open Enterprise-ready Big Data solution. HDInsight provides a platform for all of your Big Data needs including Batch, Interactive, No SQL and Streaming. It also comes with a strong eco-system of tools and developer environment.

Supported cluster types include: Hadoop (Hive), HBase, Storm, Spark, Kafka, Interactive Hive (LLAP), and ML Services.

Media

More products from Microsoft

See All

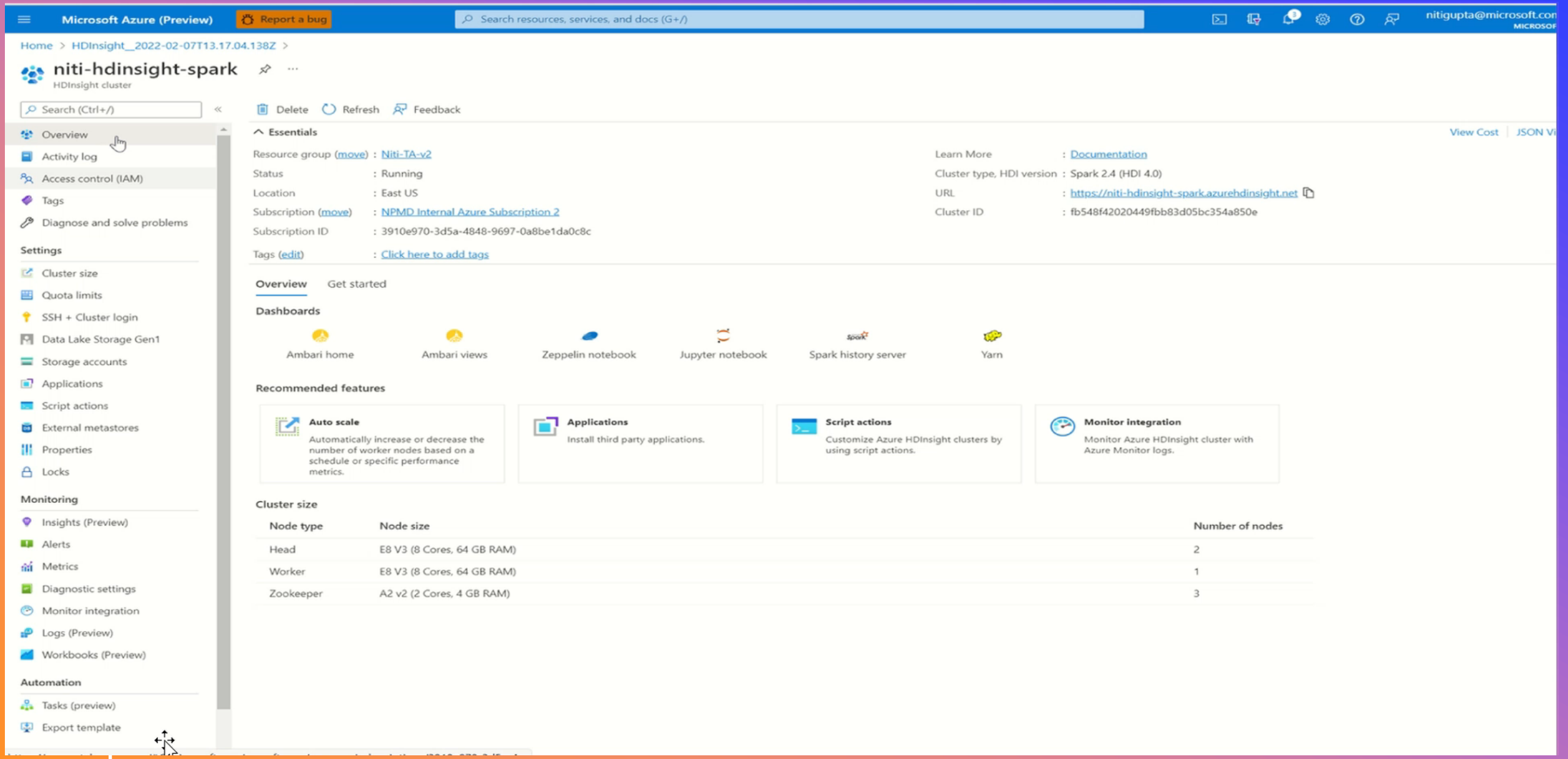
Device Update for IoT Hub
Microsoft
Azure Service
Securely and Reliably update your devices with Device Update for IoT Hub.

Front Door Standard/Premium
Microsoft
Azure Service
Azure Front Door Standard/Premium (Preview) is security led, modern cloud CDN that provides static and

Azure VMware Solution
Microsoft
Azure Service
Azure VMware Solution (AVS) combines the VMware Software Defined Data Center (SDDC) with

API App
Microsoft
Azure Service
scalable RESTful API with enterprise grade security, simple access control and auto SDK generation

DASHBOARD VIEW



KEY TAKEAWAYS

- Data is growing tremendously
- Hadoop is an efficient distributed framework for parallelly processing
- HDInsight is a highly scalable and coherent cloud solution for bid data processing and storage
- Seamless setup for azure HDInsight and numerous integration capabilities for moving from on-premise to cloud solution.

@nitigupta_16

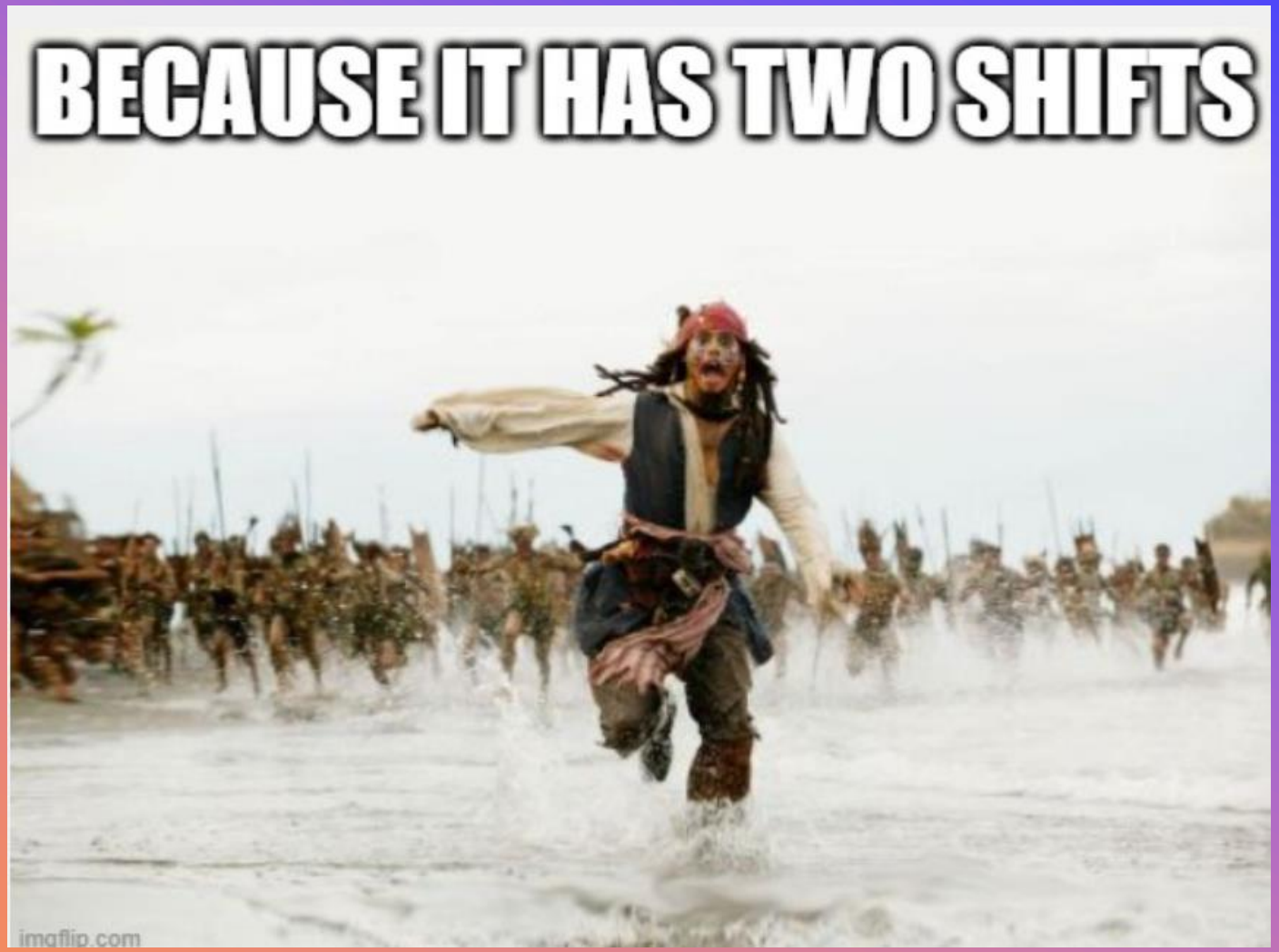
nitigupta@microsoft.com

https://www.linkedin.com/in/niti-gupta-33200086

Slides at: https://github.com/nitigupta16/talks

Why don’t keyboards sleep?

@nitigupta_16


BECAUSE IT HAS TWO SHIFTS
imgflip.com

Q & A

Thank you