



Python For Databricks

Introduction



By
Harishkumar
Coddington College
SYSTEMITY IT LABS

About Me

- ❖ Harishkumar CH
- ❖ Senior Tech Lead
- ❖ 10 Years in Solution development and Architecture design
- ❖ Social media:
 - Facebook: /harishkumarreddy.cherla
 - FB Page: /sysentity
 - LinkedIn: /in/harishkumarreddy
 - Whatsapp: +91 7997156656





Agenda

Level-1:

1. Big Data
2. Evaluation of Spark and it's features
3. Databricks Tool and Python use case
4. Python Setup and configure
5. First Programm
6. Core Python
7. Advanced Python
8. Algorithms
9. Data Structures
10. Data Processing packages(Numpy, Pandas, SciPy)

Level-2:

1. Databricks UI Overview
2. Creating and managing Clusters and Jobs
3. Playing with Notebooks
4. RDD, Dataframe, Dataset
5. DBFS and Datatables
6. Mounting / Connecting with outside Data sources
7. PySpark
8. Playing with realtime data with real time scenarios

Duration: 45 to 60 days



Requirements and Results

PreRequisites

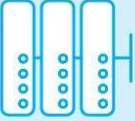





- ❖ Basic system knowledge
- ❖ Understanding of business process
- ❖ Basic Domain knowledge
- ❖ No requirement of programming

Gains

- ❖ Real Time experience
- ❖ Profitionality in Python
- ❖ Confidence to crack the interview

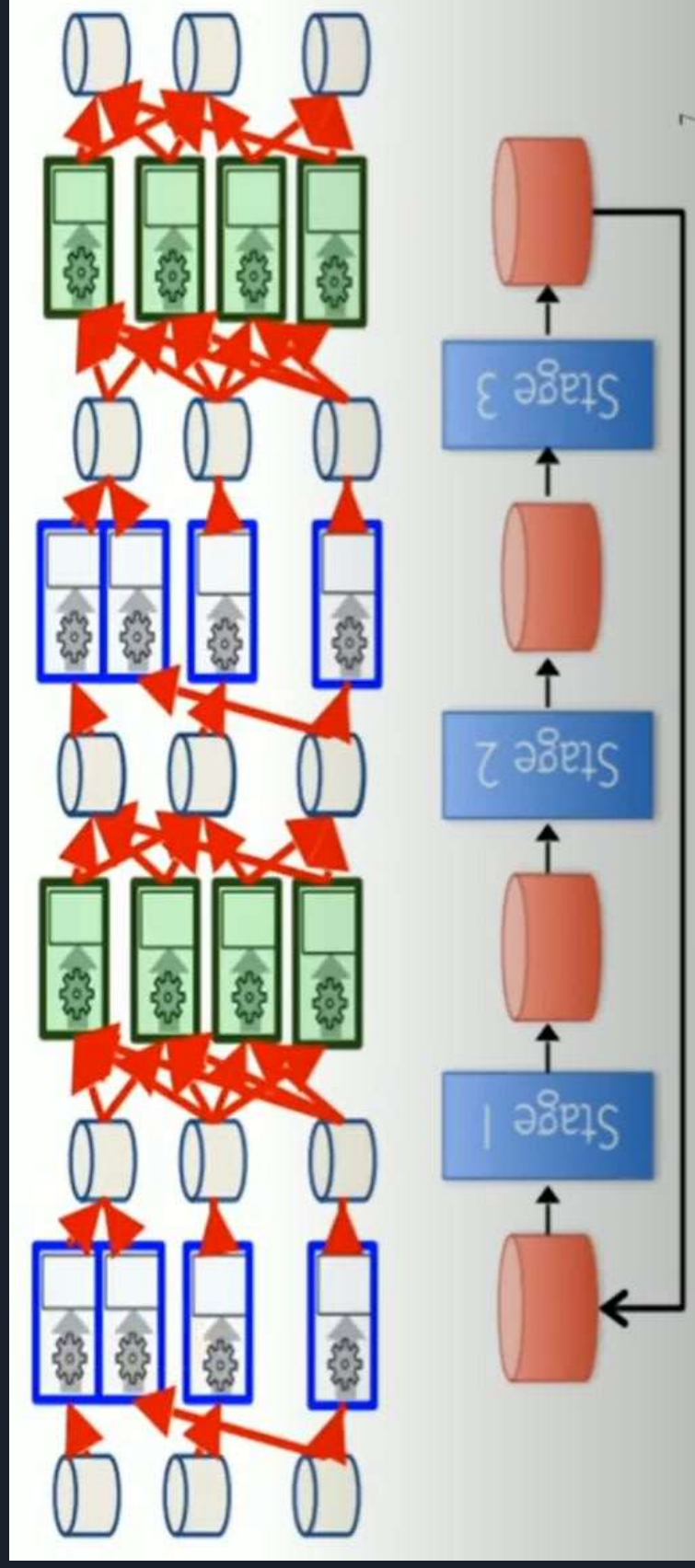
Big data

Big data is a term that describes the large volume of data – both structured and unstructured – that inundates a business on a day-to-day basis.

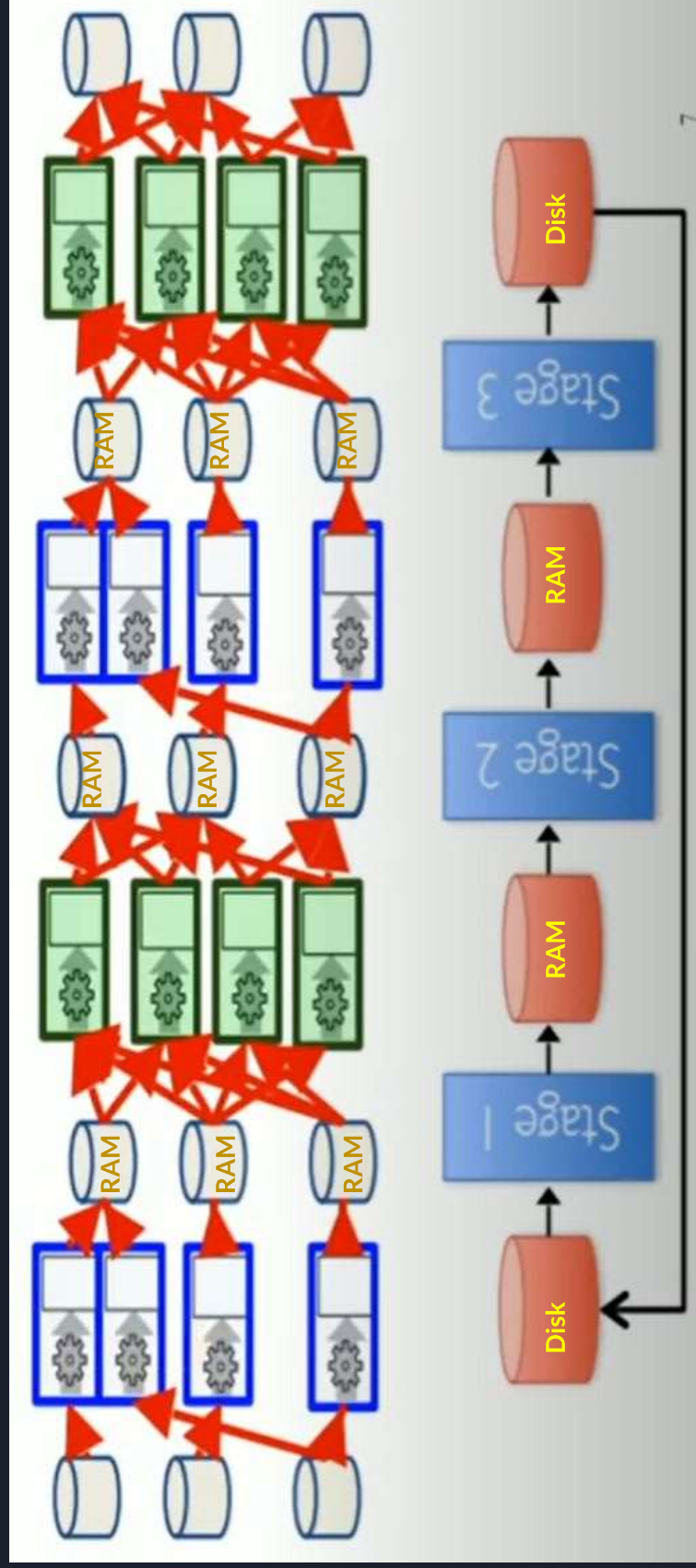
VOLUME	VARIETY	VELOCITY	VERACITY	VALUE	VARIABILITY
The amount of data from myriad sources 	The types of data: structured, semi-structured, unstructured 	The speed at which big data is generated 	The degree to which big data can be trusted 	The business value of the data collected 	The ways in which the big data can be used and formatted 



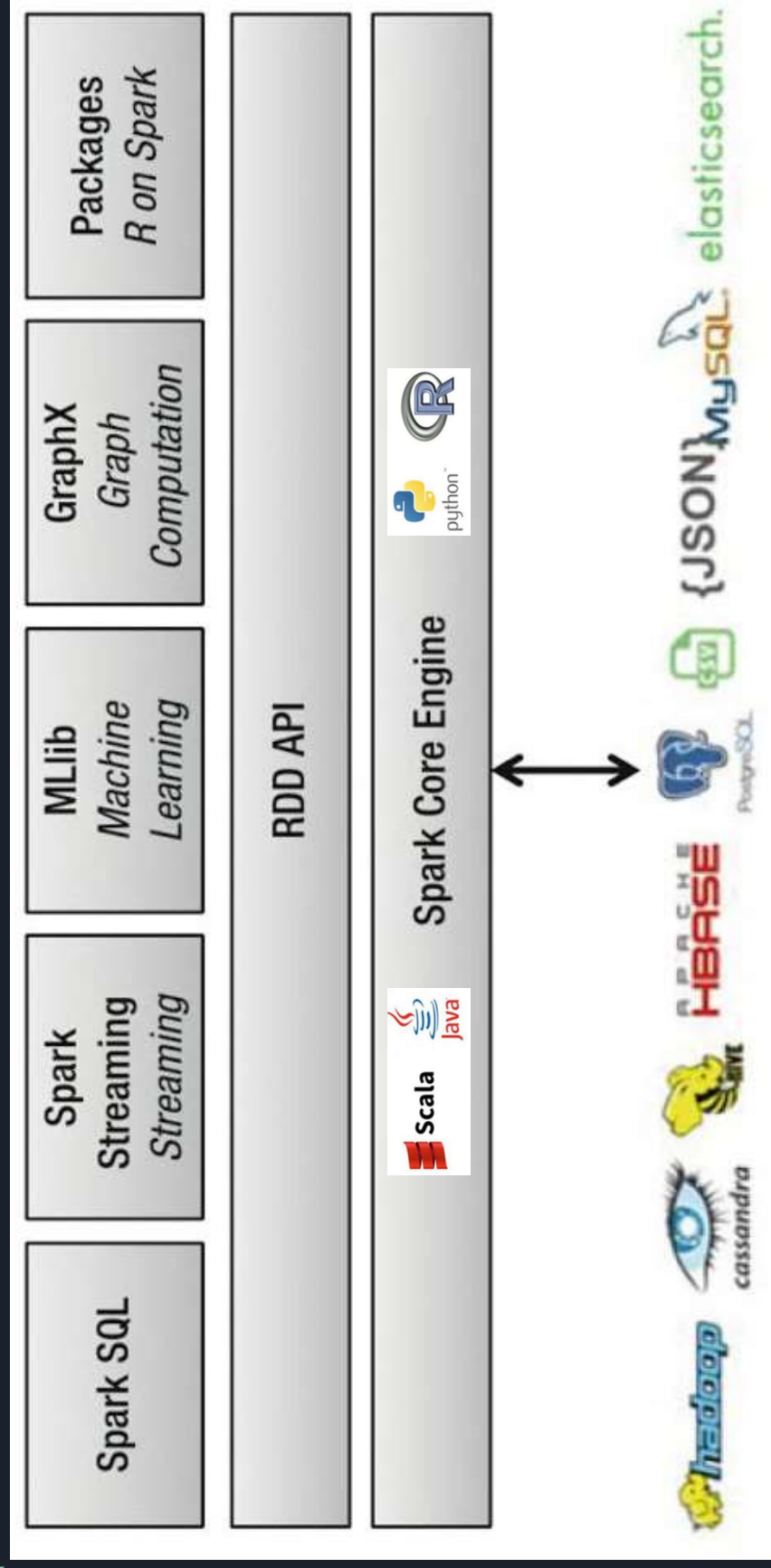
I/O Problem in Hadoop



Spark Solution for I/O Problem

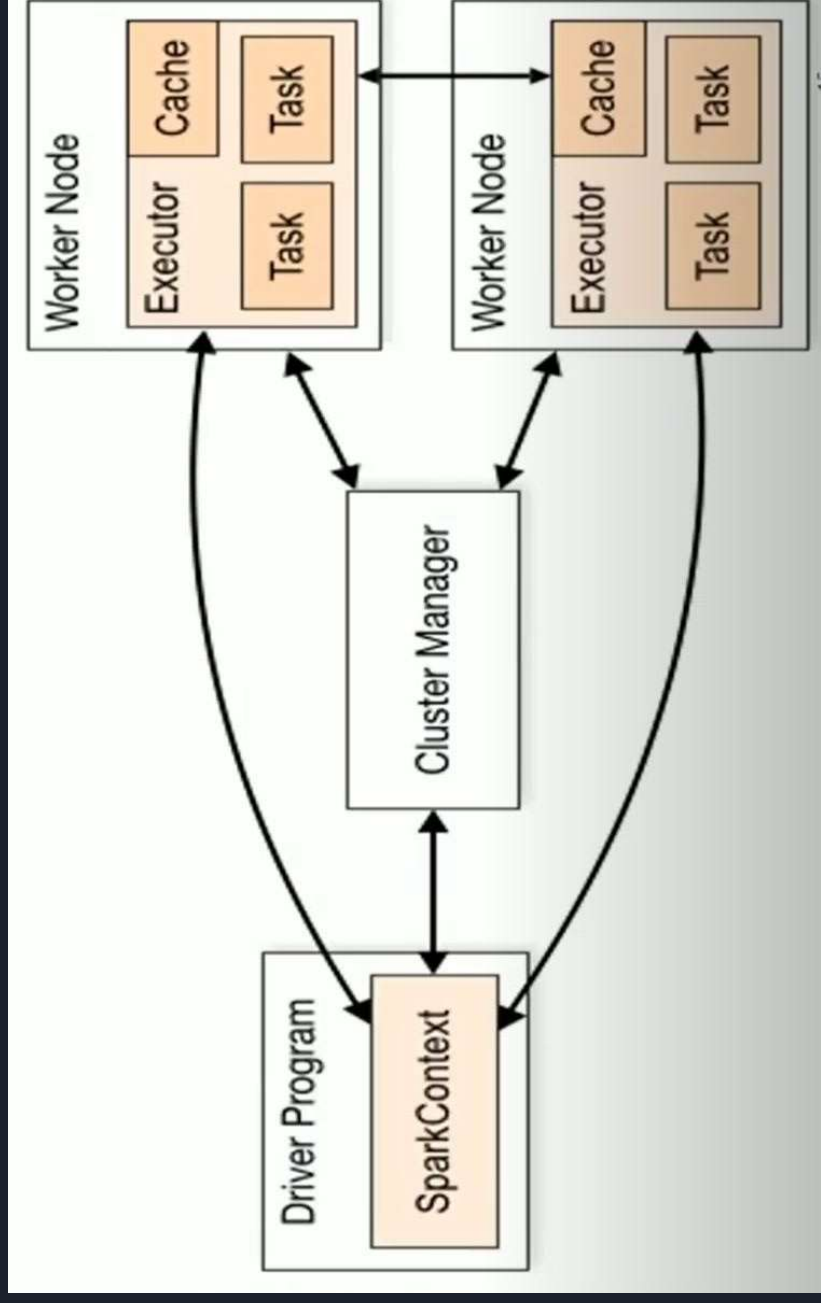


Spark for Big Data



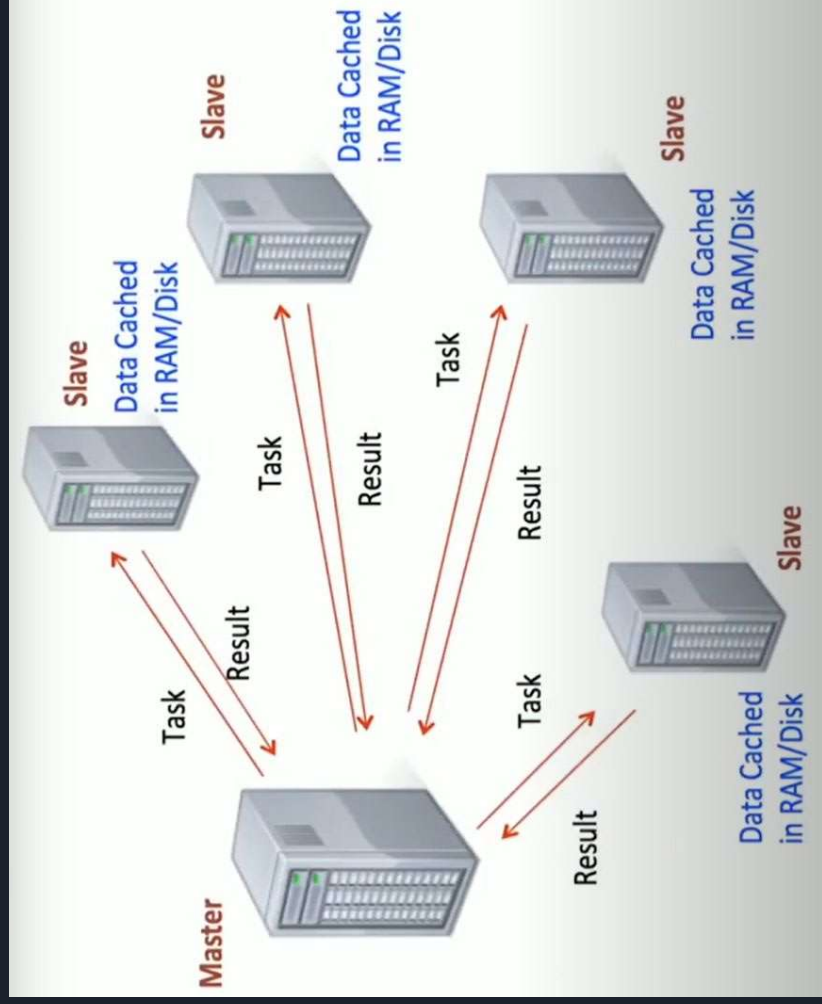


Execution in Spark





Master - Slave shakehand

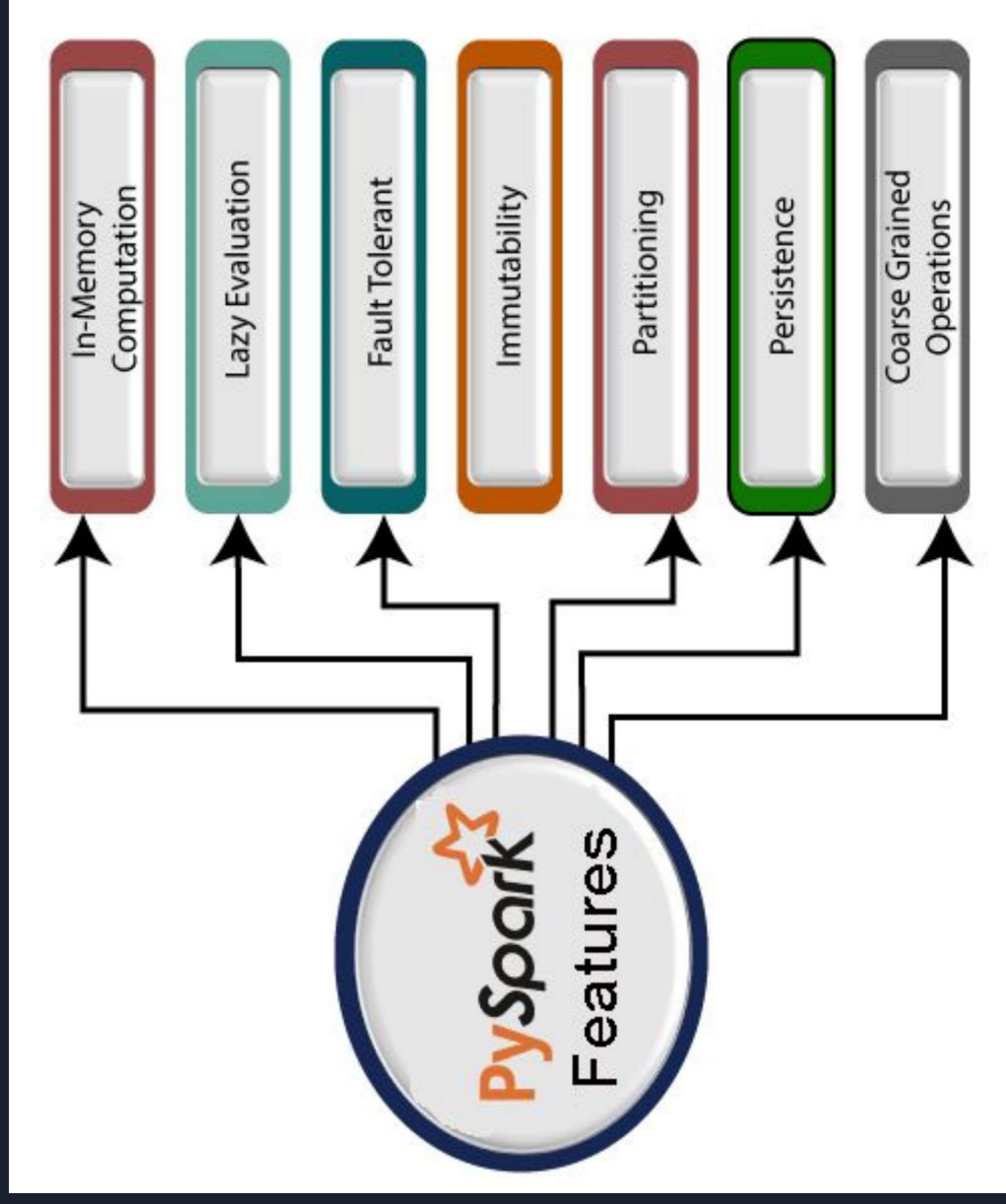





DW vs MR vs Spark










	Data Warehouse	Hadoop M/R	APACHE Spark™
Separate Compute & Storage	✗	✓	✓
More than SQL (i.e ML)	✗	✓	✓
Open Source at Scale	✗	✓	✓
SQL & Optimization	✓	✗	✓
Data Model & Catalog	✓	✗	✓
ACID Transactions	✓	✗	✓
			3.0
			DELTA LAKE


Python with Spark




Databricks










Explore the Quickstart Tutorial

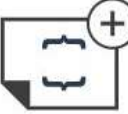
Spin up a cluster, run queries on preloaded data, and display results in 5 minutes.

Drop files or click to browse



Import & Explore Data

Quickly import data, preview its schema, create a table, and query it in a notebook.



Create a Blank Notebook

Create a notebook to start querying, visualizing, and modeling your data.

New Notebook

Create Table

New Cluster

New Job

New MLflow Experiment

Import Library

Read Documentation

mnist-pytorch

test_0504

Quickstart Notebook


mnist-tensorflow-to-tfrecords

What's new in v3.45

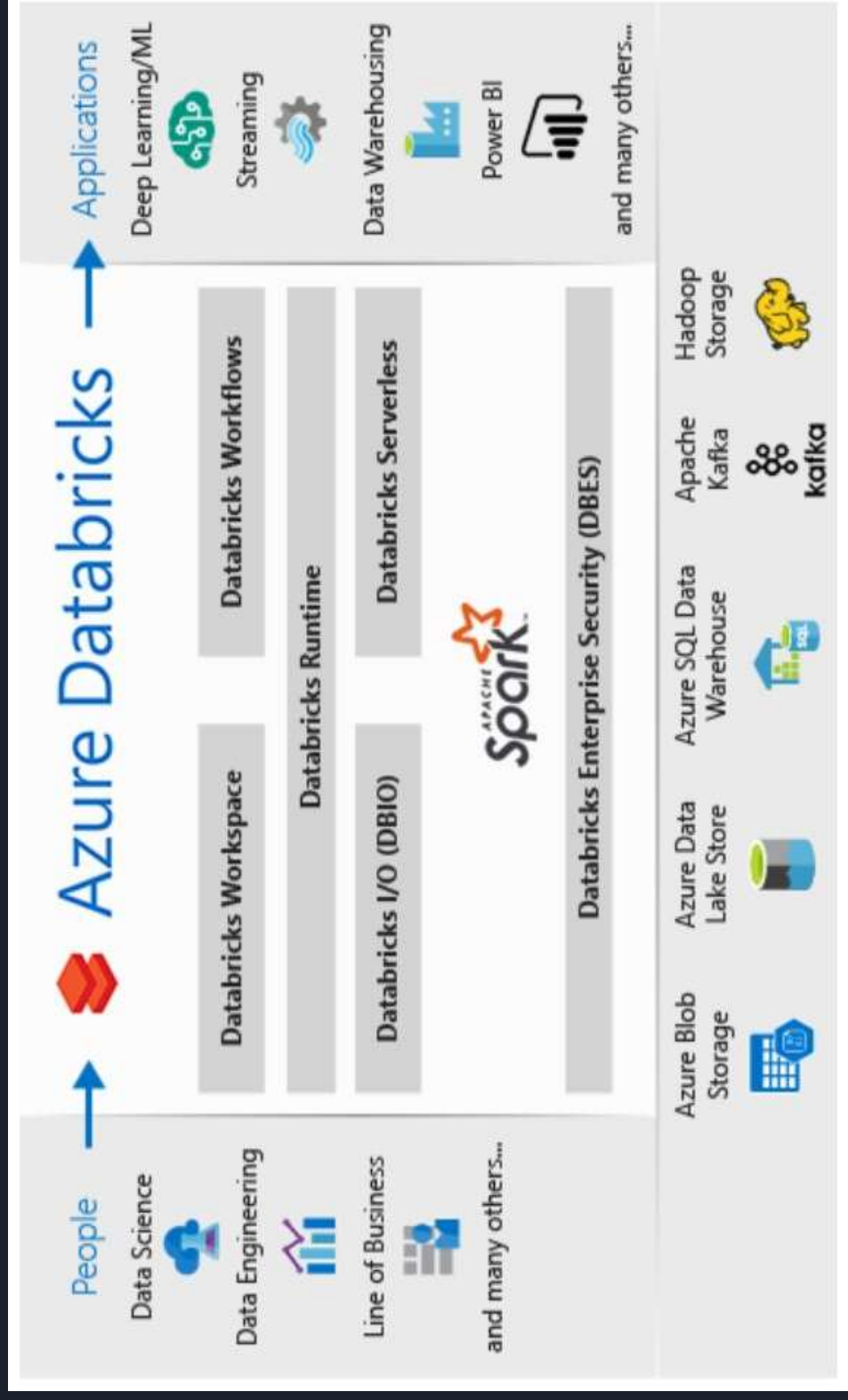
View latest release notes

?

E2



Databricks Workflow





What Next

1. Setup and configure Python
2. Code editors & IDEs
3. Run the first program
4. Basic concepts of programming
5. Core Python

