**Minimum Tool & Environment Setup**

**1. Core Tools**

| **Tool** | **Version** | **Purpose** |
| --- | --- | --- |
| **Apache Spark** | 3.3+ | Run Spark jobs (local or distributed) |
| **Python** | 3.8+ | Use PySpark APIs |
| **Java** | 8 or 11 | Spark runtime requirement |

**2. Cluster & Runtime**

| **Option** | **Setup Type** | **Usage** |
| --- | --- | --- |
| **Local Spark (Standalone)** | Laptop | For basic testing, DAG, RDD vs DF, cache labs |
| **Databricks Community Edition** | Managed Cloud | Great for DAG, UI, UDF tuning, Azure integration |
| **Azure HDInsight / Synapse Spark Pool** | Azure-based Spark | Needed for Blob/ADLS read experiments |

**3. IDE & Notebook**

| **Tool** | **Usage** |
| --- | --- |
| **Jupyter Notebook** | Best for interactive Spark experiments using PySpark |
| **VS Code with PySpark Plugin** | Great for local dev with IntelliSense |
| **Databricks Notebook** | Cloud-native, fully-featured Spark IDE |

**4. Storage & Data Access**

| **Storage** | **Needed for** |
| --- | --- |
| **Local CSV/JSON/Parquet files** | Practice for RDD vs DF, cache, partitioning |
| **Azure Blob Storage or ADLS Gen2** | Needed for external source read labs |
| **Mounting via wasbs:// or abfss://** | For ADLS/Blob access from Spark |

Spark must be configured with proper Hadoop and Azure connectors for ADLS:

bash

--packages org.apache.hadoop:hadoop-azure:3.3.1,com.microsoft.azure:azure-storage:8.6.6

**5. Monitoring Tools**

| **Tool** | **Purpose** |
| --- | --- |
| **Spark UI (**[**http://localhost:4040**](http://localhost:4040)**)** | DAG analysis, shuffle, cache visibility |
| **Spark History Server** *(Optional)* | Analyze saved Spark event logs |
| **Ganglia / Prometheus** *(Optional)* | Cluster-level monitoring (advanced) |

**6. Dependencies & Libraries**

| **Library** | **Why** |
| --- | --- |
| pandas | For pandas\_udf |
| pyarrow | Needed for Arrow-based serialization |
| openpyxl | Optional for exporting reports |
| findspark | Helpful in Jupyter notebooks |

Install sample:

bash

pip install pandas pyarrow findspark openpyxl