



Street Smart Machine Learning Part I: Breadth with Tableau, R & Spark

Drew Minkin
Lead Instructor and Curriculum Dev
Data Science Immersive Program
Nov 4 2022



Topics Overview

1. Tableau & R Integration
2. Spark and Big Data
3. Spark & R Integration
4. Street Smart Machine Learning

Who's Drew?

Previous Incarnations

SQL Support/Consulting 2000-06



Data Scientist 2010 – Now



4.5 Startups 2007-19



Analytics Architect/Dev



Entrepreneur



Current Incarnation

MCT, Data Science Instructor

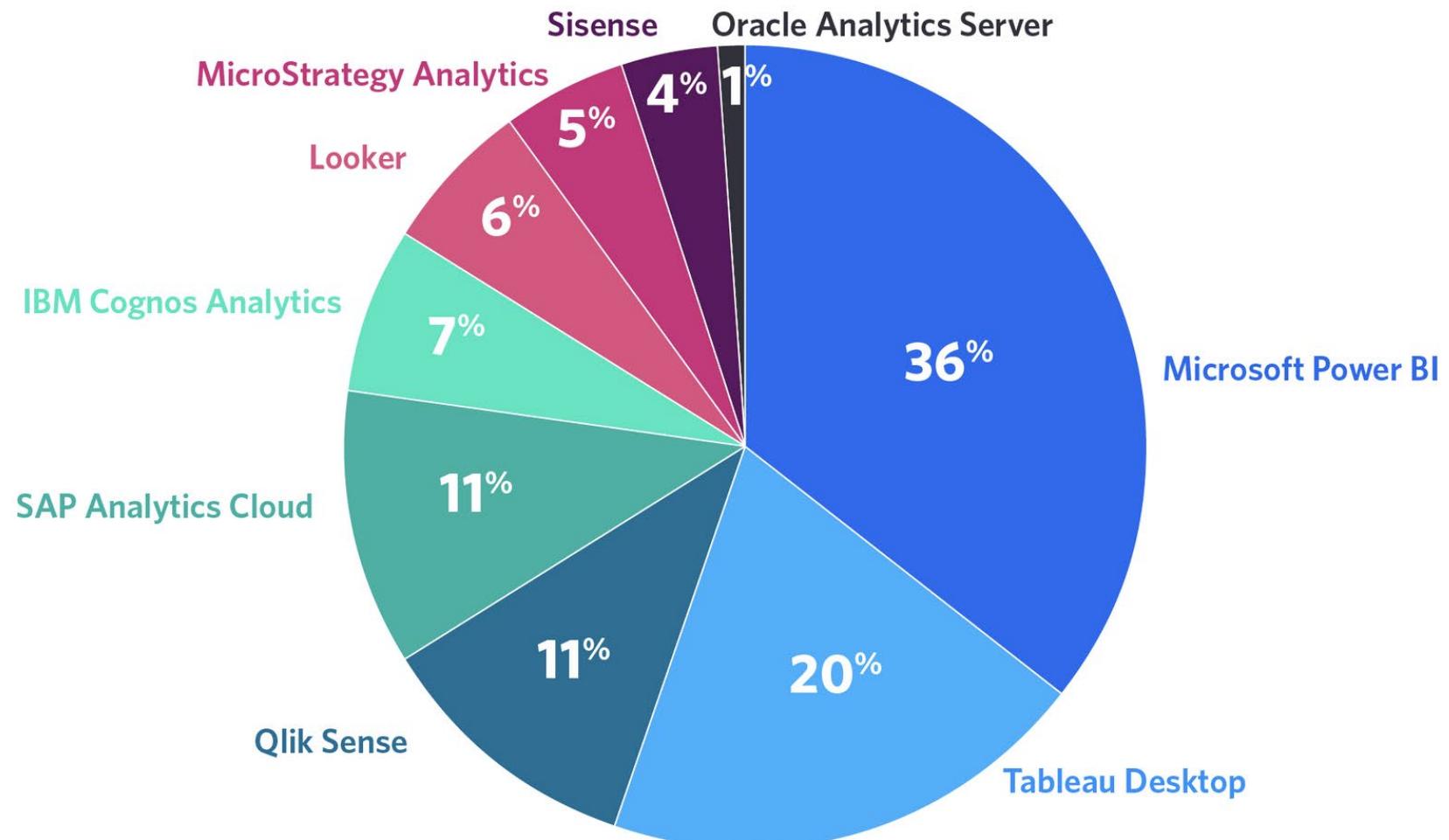




Tableau & R Integration



Tableau and Self-Service Business Analytics



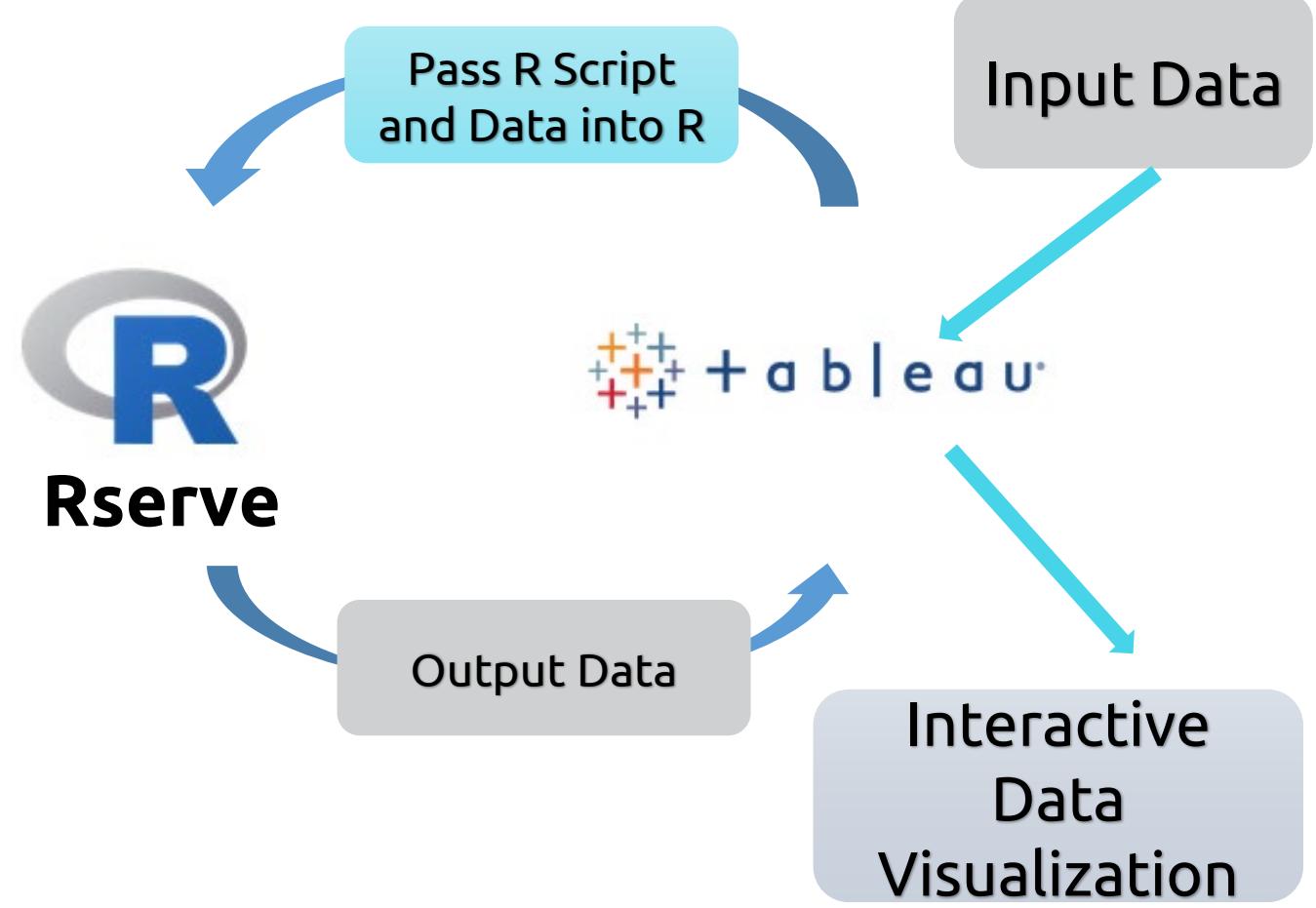
Source: TrustRadius platform data collected in June 2021

© TrustRadius

Introduction to Tableau and R Integration

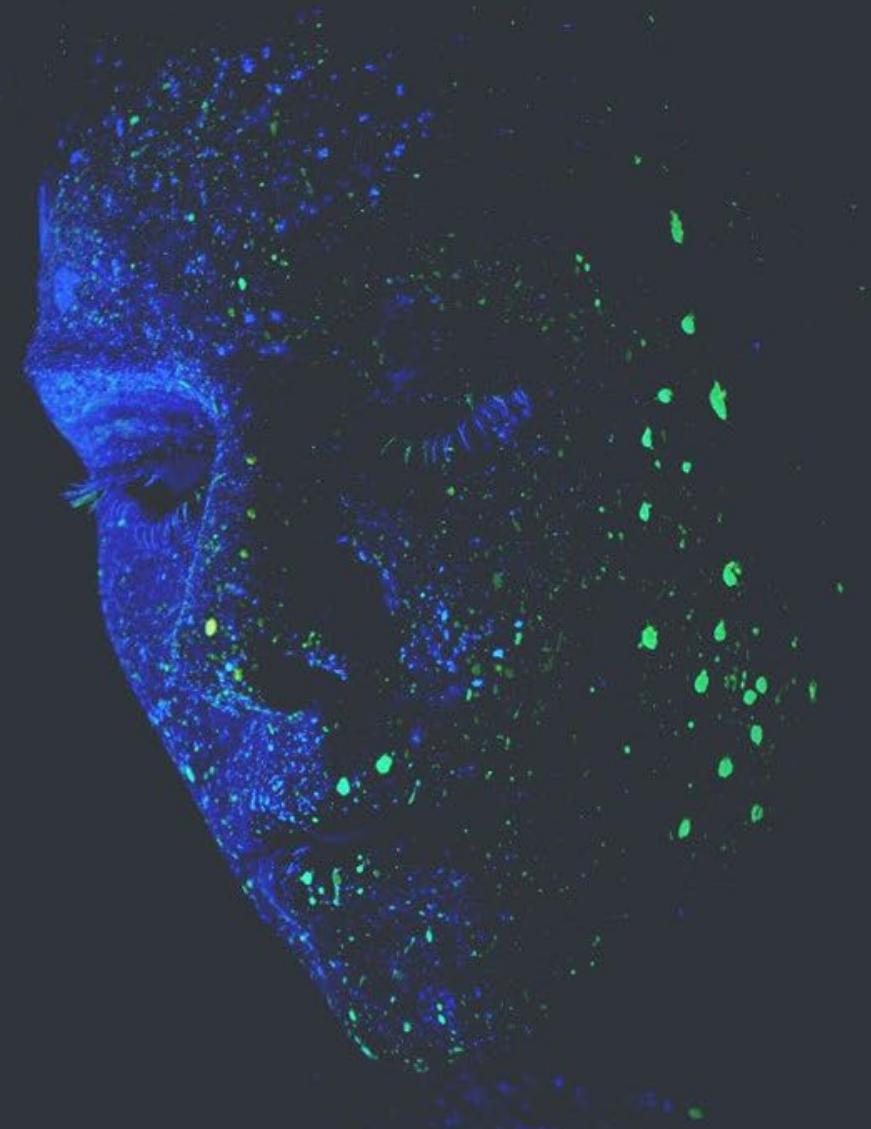


- Use one of the four functions
 - SCRIPT_BOOL:** Return a Boolean
 - SCRIPT_INT:** Return an Integer
 - SCRIPT_REAL:** Return a Real
 - SCRIPT_STR:** Return a String



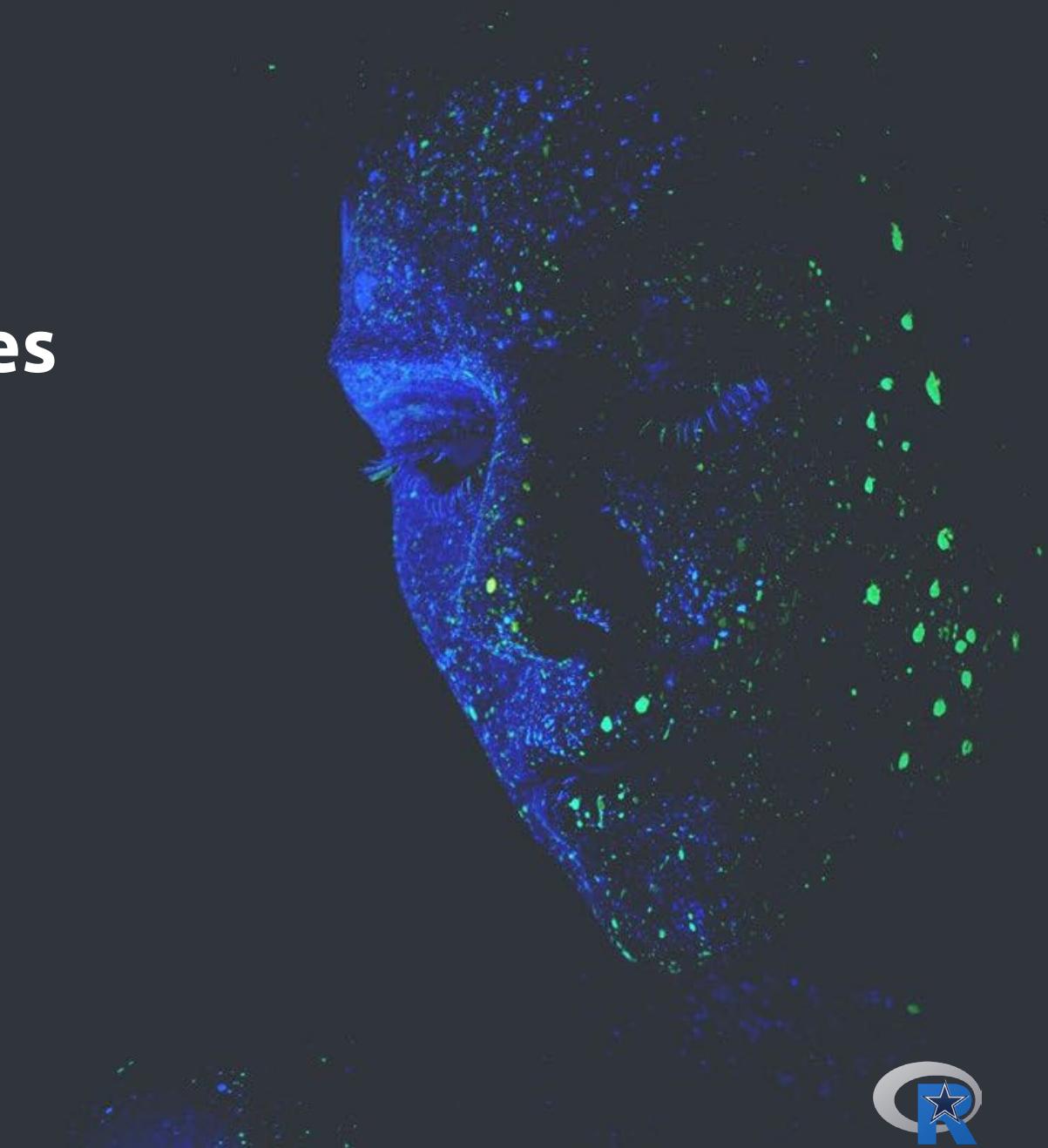


Demo: Tableau & R Integration





A Tale of Two Technologies And the Future of Data



Why Care about Big Data



THIN THE LINES
BETWEEN
OPERATIONAL AND
ANALYTICAL



DECENTRALIZED
WORKSPACES
CENTRALIZED STORAGE



COMPUTE IS
PAY AS YOU GO

Data Structures Beyond the Relational RDBMS



Key Value



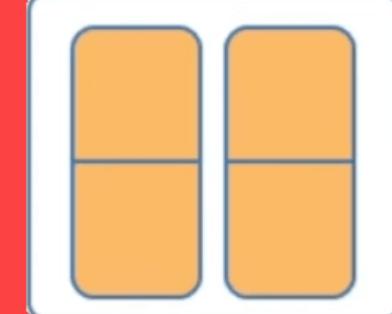
Example:
Riak, Tokyo Cabinet, Redis
server, Memcached,
Scalarmis

Document-Based



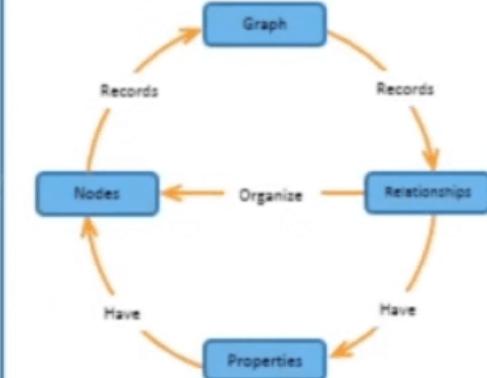
Example:
MongoDB, CouchDB,
OrientDB, RavenDB

Column-Based



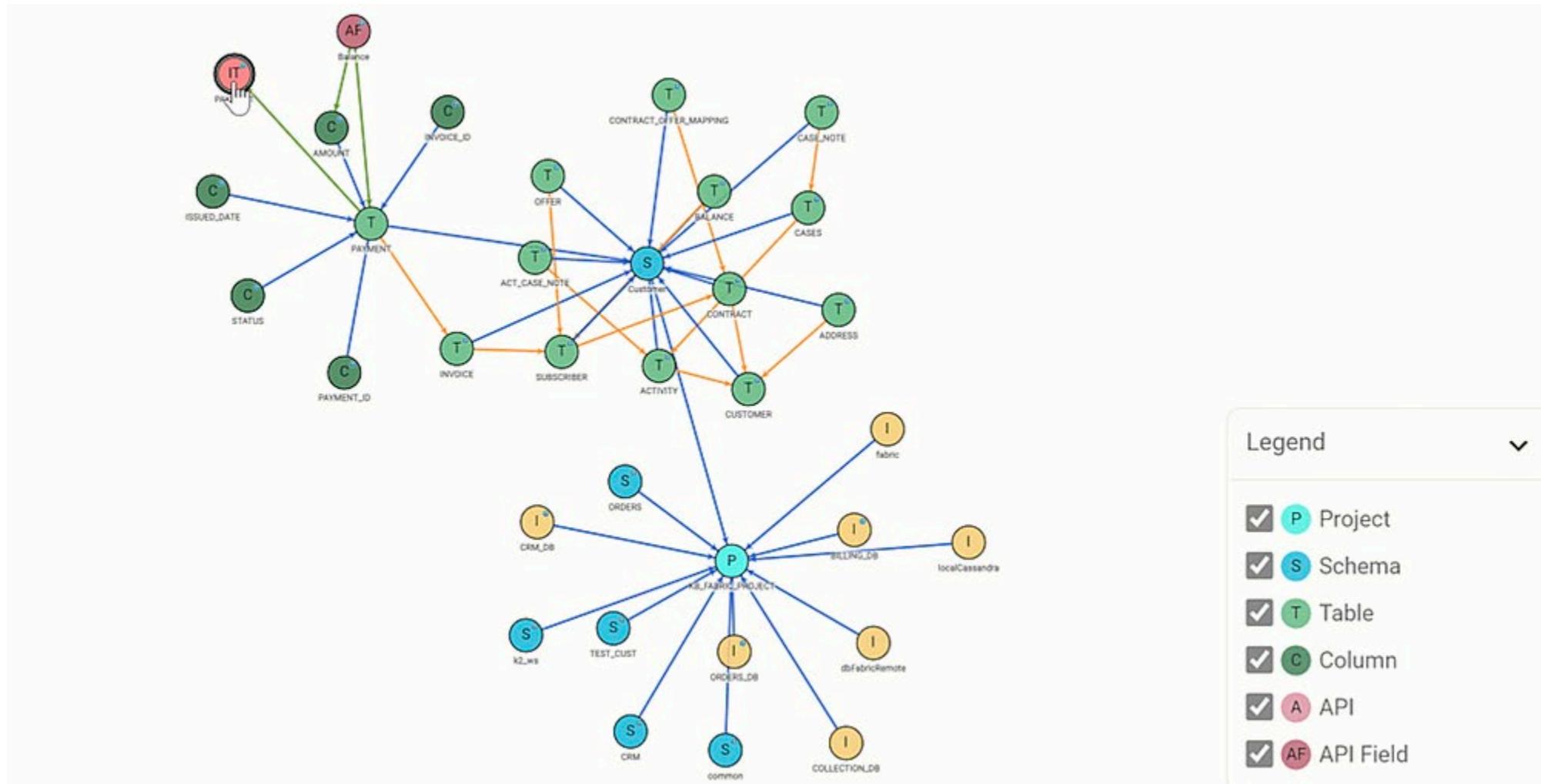
Example:
BigTable, Cassandra,
Hbase,
Hypertable

Graph-Based

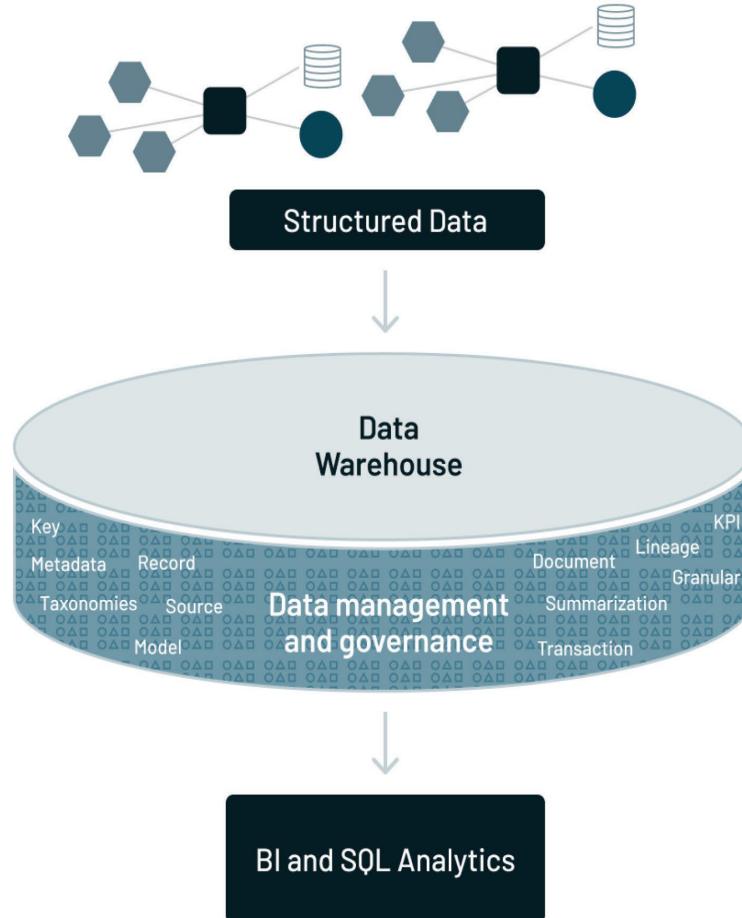


Example:
Neo4J, InfoGrid, Infinite
Graph, Flock DB

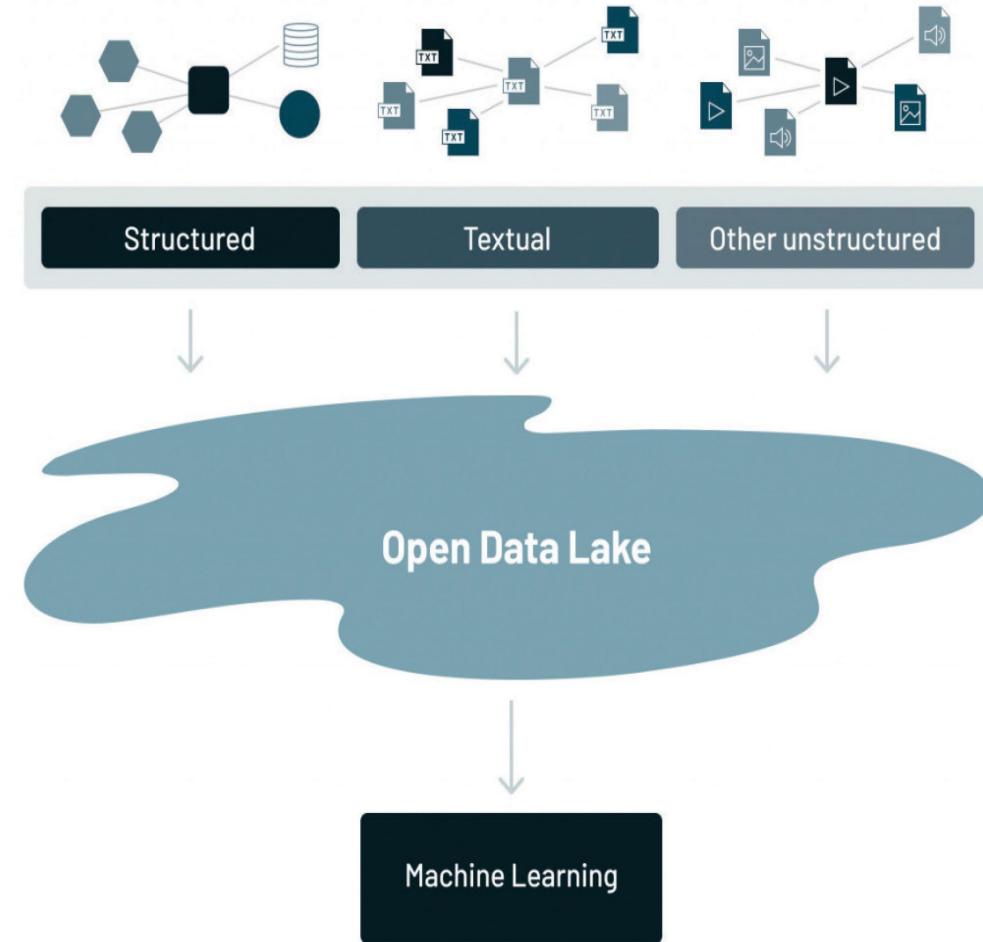
From data to information: the Knowledge Graph



Comparison of Technologies



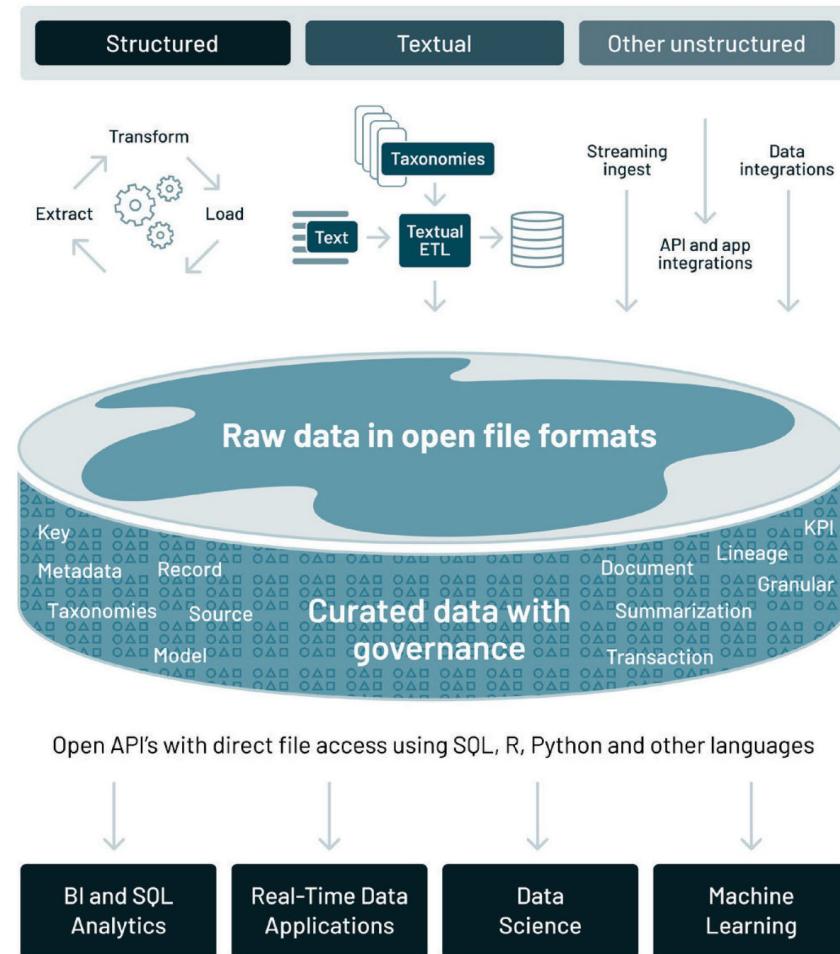
Comparison of Technologies



Comparison of Technologies



Data Lakehouse

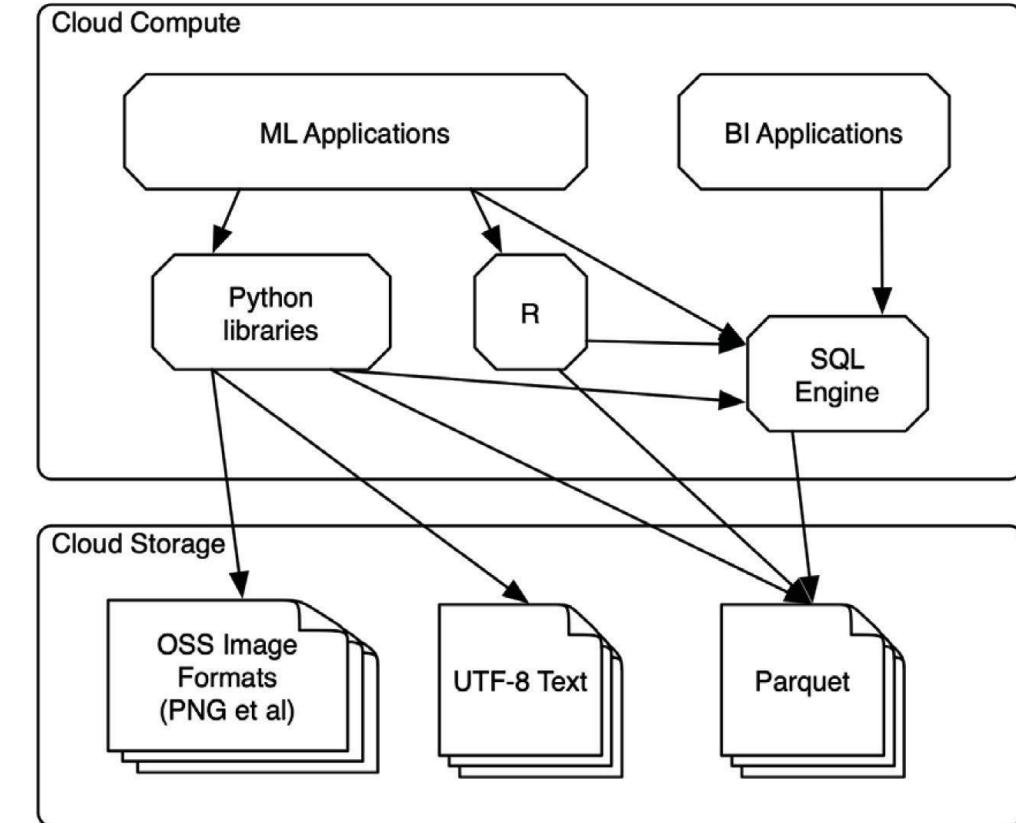
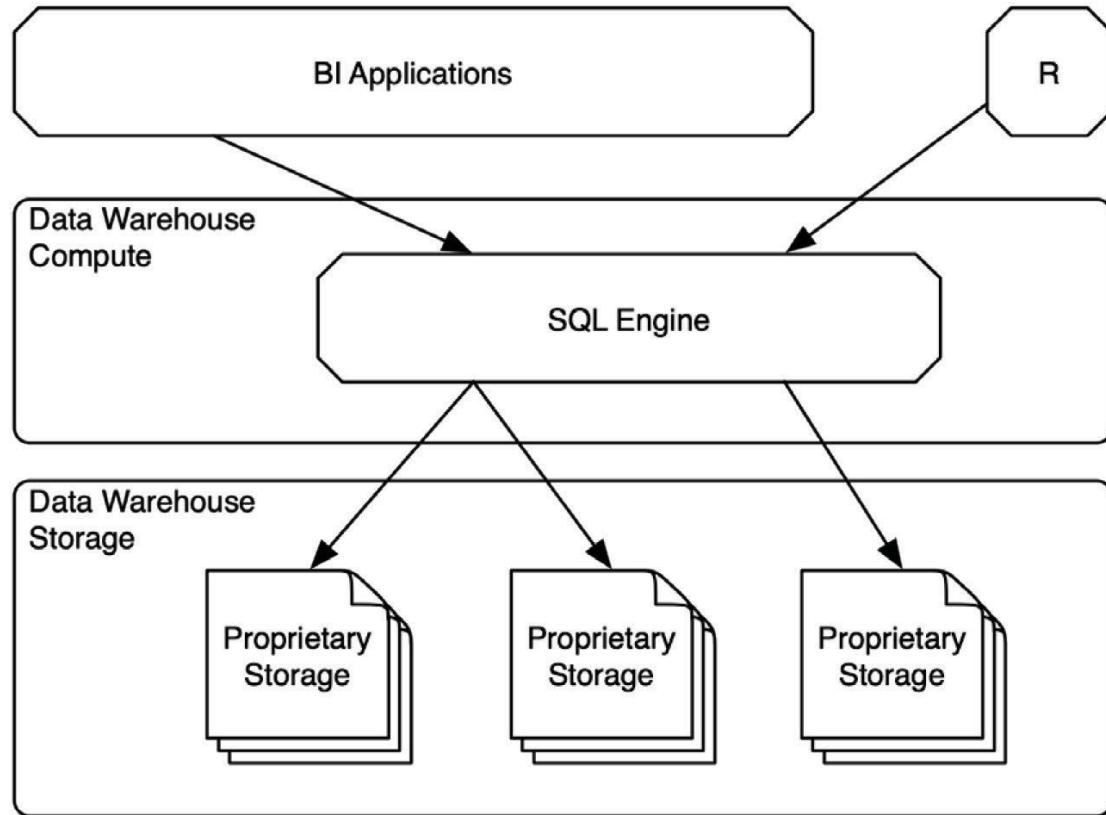


Comparison of Technologies

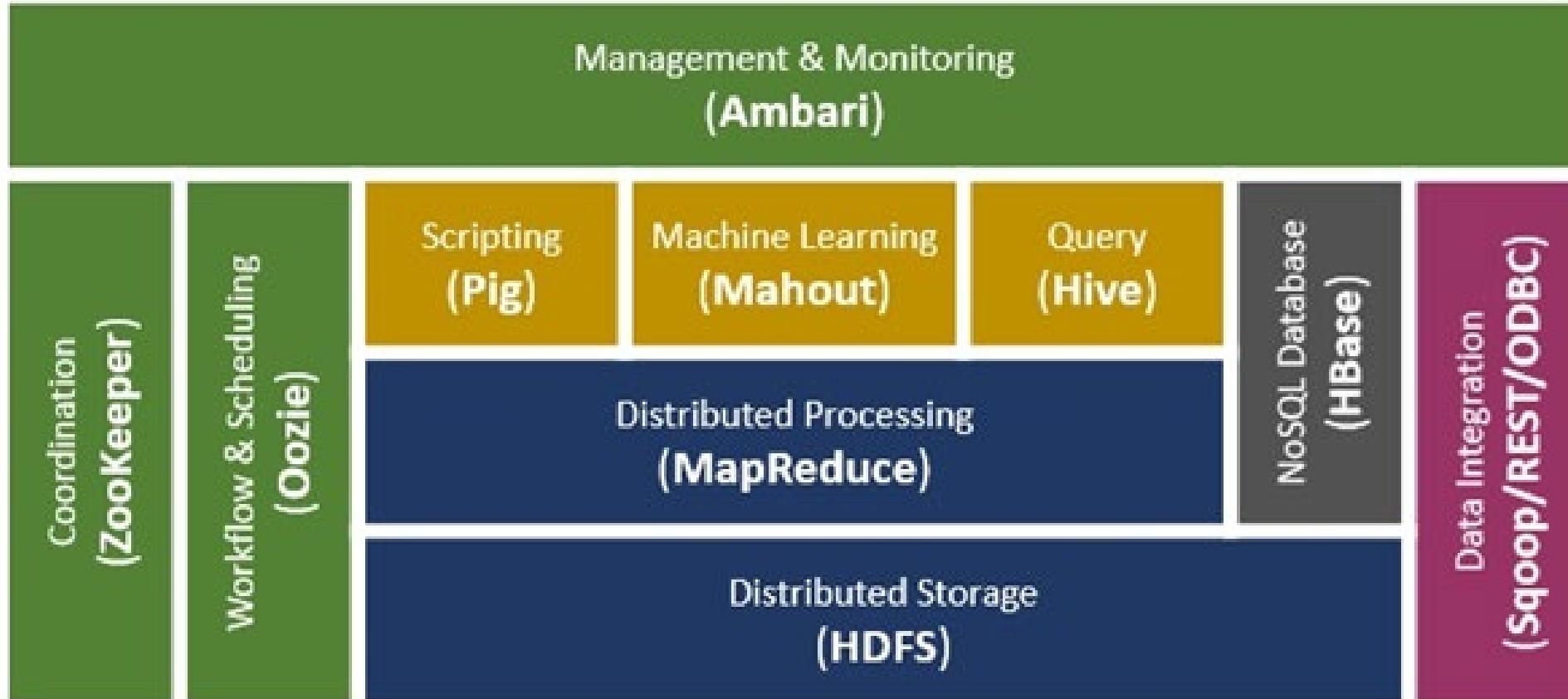


	Data warehouse	Data lake	Data lakehouse
Data format	Closed, proprietary format	Open format	Open format
Types of data	Structured data, with limited support for semi-structured data	All types: Structured data, semi-structured data, textual data, unstructured (raw) data	All types: Structured data, semi-structured data, textual data, unstructured (raw) data
Data access	SQL-only	Open APIs for direct access to files with SQL, R, Python, and other languages	Open APIs for direct access to files with SQL, R, Python, and other languages
Reliability	High quality, reliable data with ACID transactions	Low quality, data swamp	High quality, reliable data with ACID transactions
Governance/security	Fine-grained security and governance for row/columnar level for tables	Poor governance as security needs to be applied to files	Fine-grained security and governance for row/columnar level for tables
Performance	High	Low	High
Scalability	Scaling becomes exponentially more expensive	Scales to hold any amount of data at low cost, regardless of type	Scales to hold any amount of data at low cost, regardless of type
Use Case Support	Limited to BI, SQL applications, and decision support	Limited to machine learning	One data architecture for BI, SQL, and machine learning

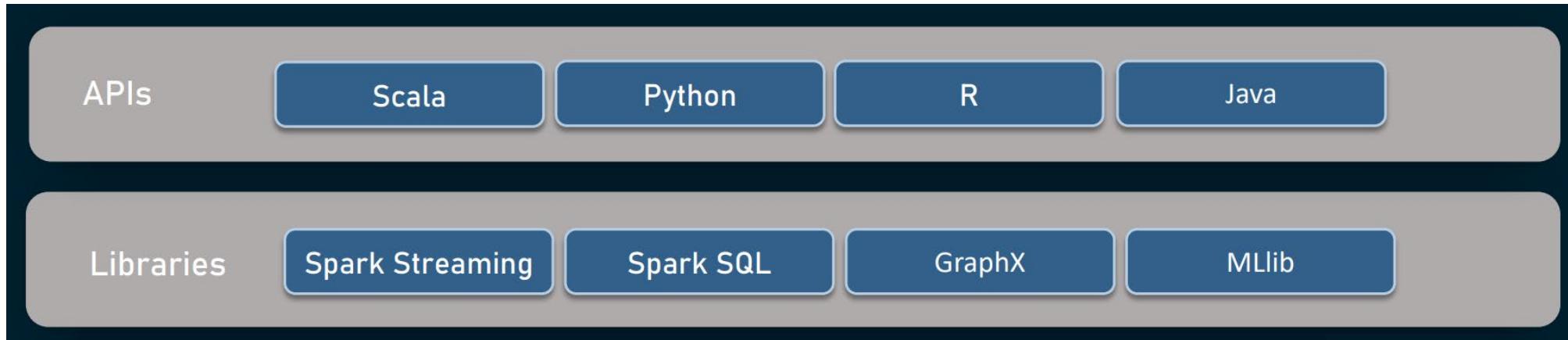
Data Warehouse vs. Data Lakehouse



Hadoop Architecture (Pre-Spark)



Spark Architecture

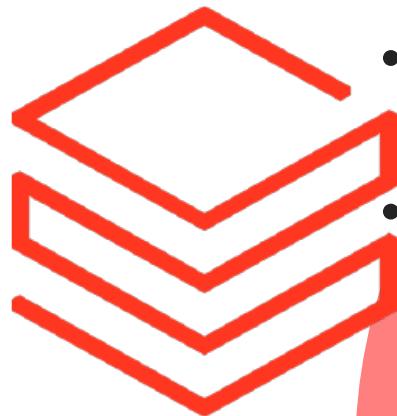


SEPARATION OF COMPUTE AND STORAGE

NATIVE JVM WITH SCALA DIALECT

SPARKR AND PYSPARK THUNKY INTEGRATIONS

Azure Synapse vs. Databricks



AZURE
DATABRICKS

- Git Support
- Deeper Spark IP
- Cross-cloud support
- Single node support
- Maturer autoscale
- Tight MLFlow integration
- Tighter Delta Lake in Spark SQL
- Spark SQL Live Tables
- Component level Azure Data Factory integration
- Notebook as Dashboard

Hive Support
Java/Scala/Python
Spark Scaleout
GPU Support
SparkML Support

- SynapseML support
- Serverless or Dedicated SQL Pool
- Data warehousing
- Data Lake Tables
- Tighter Power BI integration
- .NET language support
- Big Data Scale Model
- Scoring with Zero Data Movement
- Transact-SQL compatibility
- Azure Purview Integration
- Azure Data Factory workspace integration

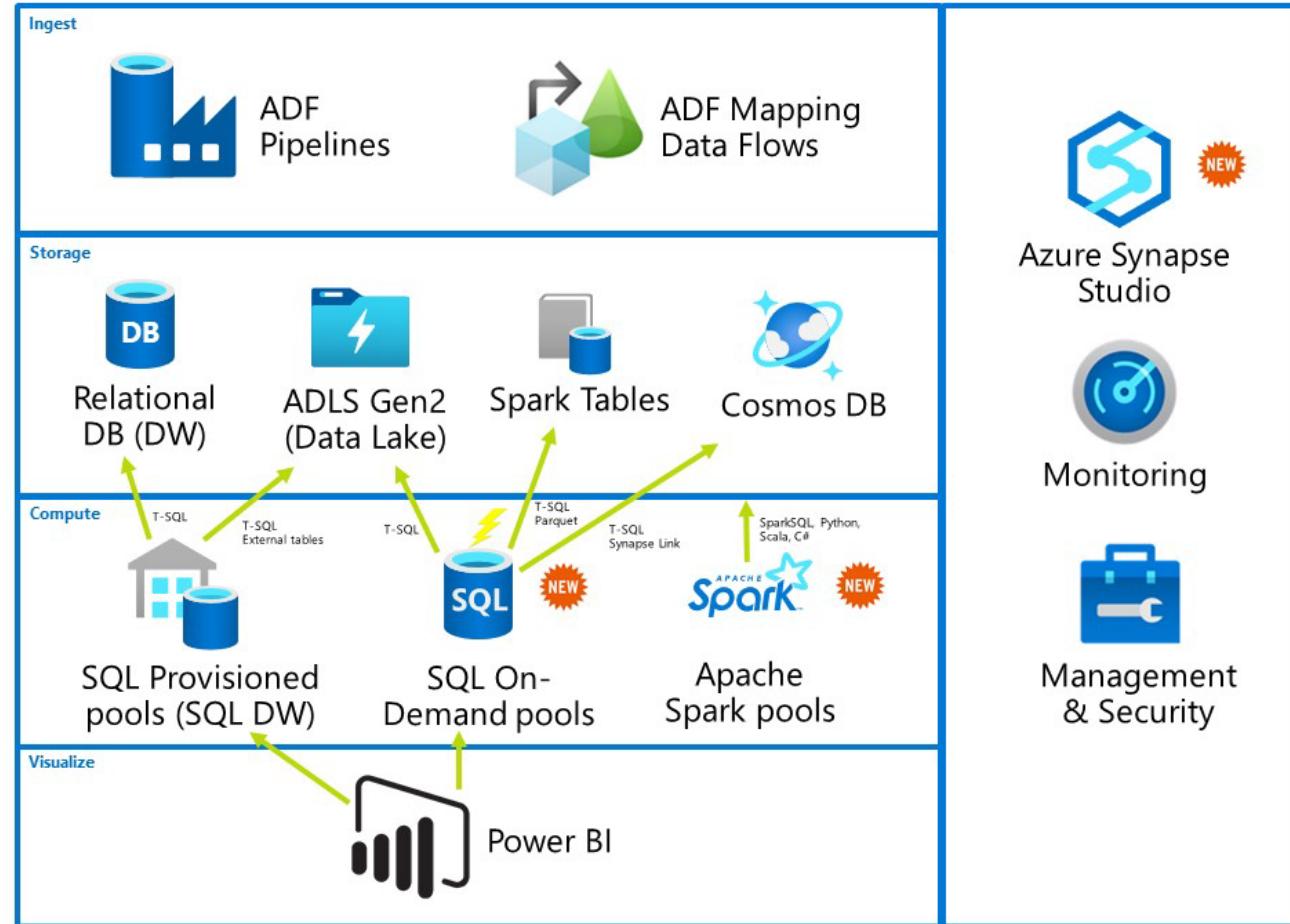


AZURE
SYNAPSE

Azure Synapse Architecture



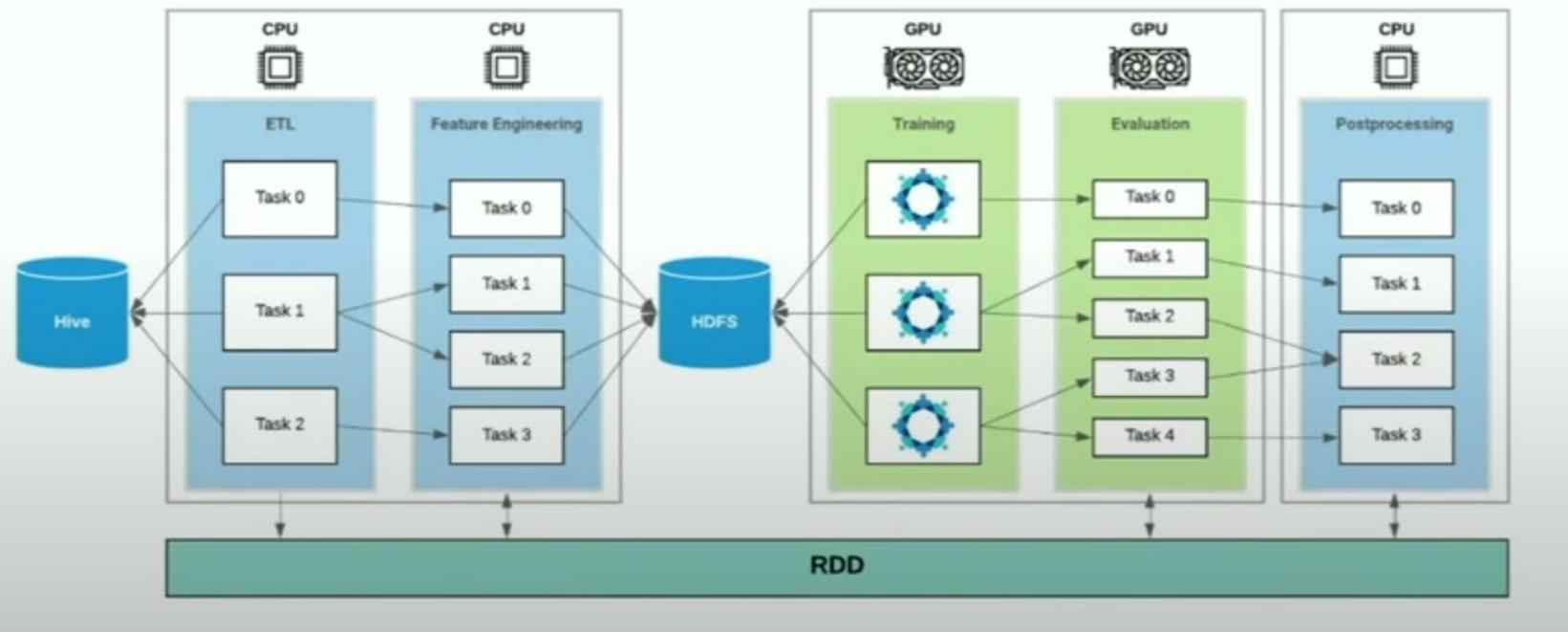
Azure Synapse
Analytics
(workspaces)



Azure Synapse Analytics – Deep Learning Benefits



- Scale-out
- Spark integration built-in
- Easier Azure DevOps & Azure Data Factory integration
- .NET language, Transact-SQL integration



<https://www.youtube.com/watch?v=jbbnZlpCu-U>

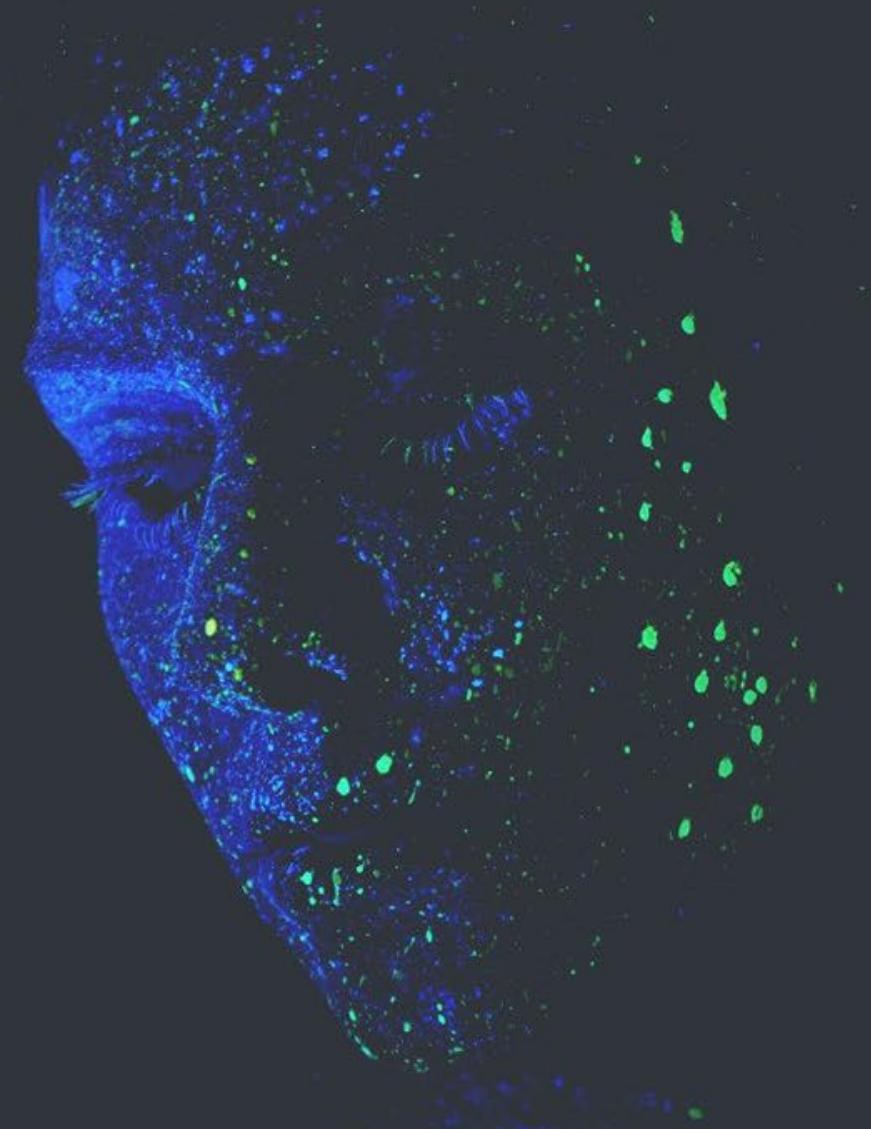


Demos: Azure Synapse & Databricks

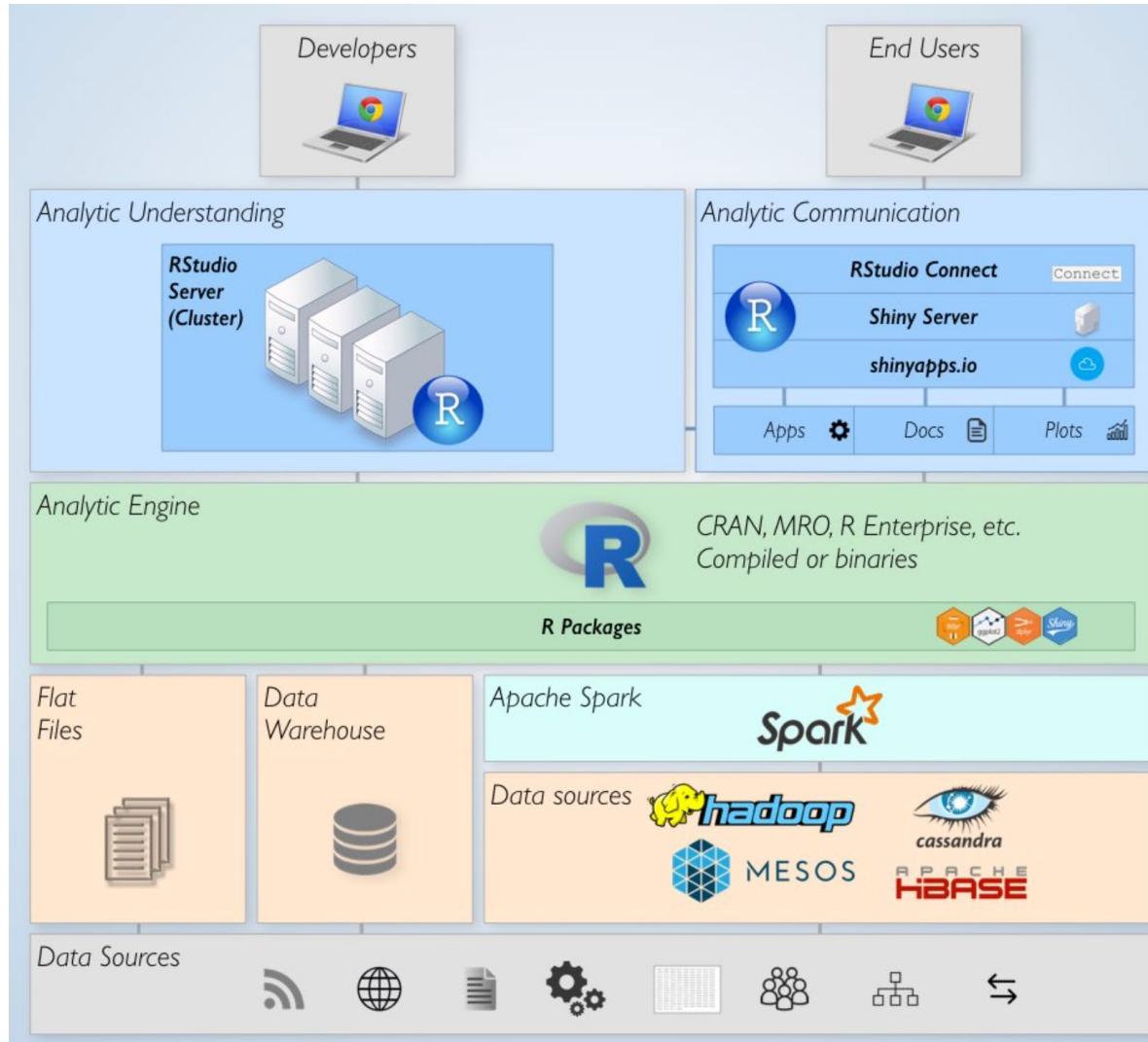




Spark & R Integration



Spark & R Integration Architectural Overview

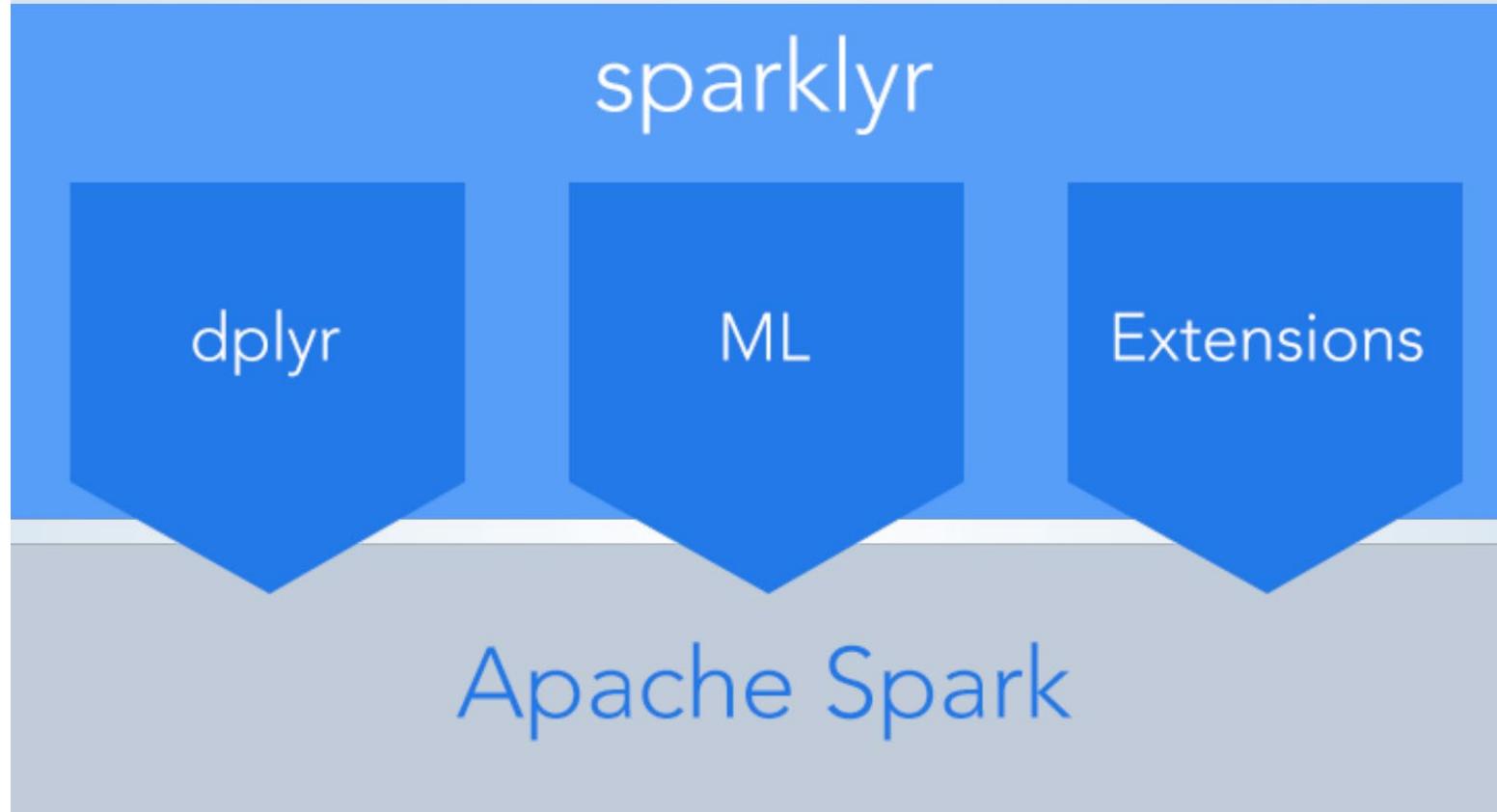


Two APIs: SparkR vs sparklyr



SparkR	sparklyr
has <i>SparkSession</i> concept	uses <i>spark_connect</i>
follows <i>R</i> syntax	benefits from <i>dplyr</i> syntax
masks some R functions	consistent naming convention
provides different <i>dplyr</i> syntax for functions	same <i>dplyr</i> functions
R to Spark DF: <i>as.DataFrame</i>	R to Spark DF: <i>copy_to</i>
for UDFs: <i>dapply</i> , <i>gapply</i> , <i>spark.apply</i>	for UDFs: <i>spark_apply</i>
ML models: <i>spark.*</i> functions, i.e. <i>spark.kmeans</i>	ML models: <i>ml_*</i> functions, i.e. <i>ml_linear_regression</i>
ML transformations: special functions, i.e. <i>add_months</i>	ML transformations: <i>ft_*</i> functions, i.e. <i>ft_string_indexer</i>
ML: no model validators nor evaluators	ML: model validators and evaluators
List of available functions	List of available functions

The winner and new champion...



Advanced Analytics – Technical Debt

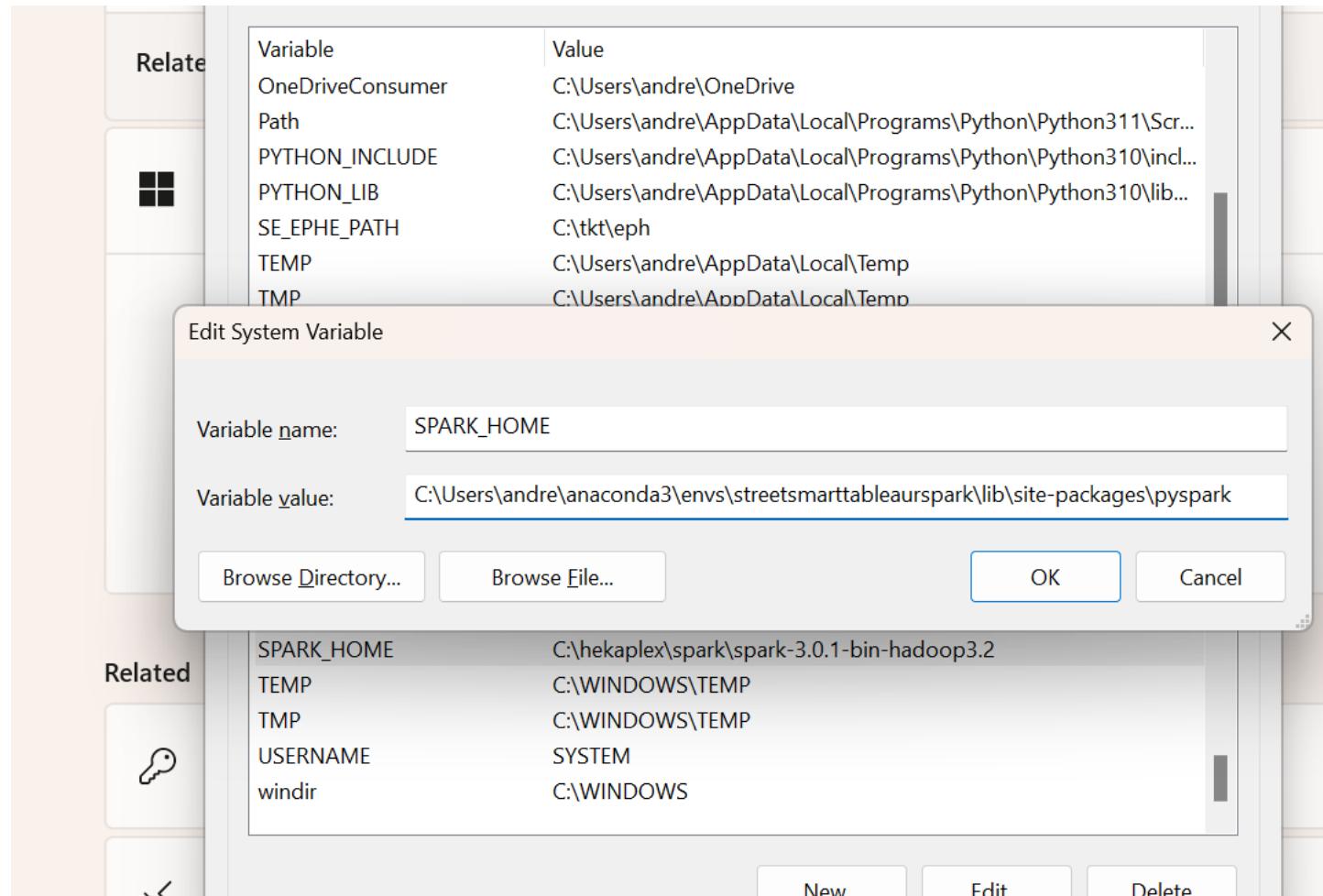
A Proportional Measure of Estimated Effort in the Entire Machine Learning Lifecycle



The screenshot shows the Microsoft Azure Databricks User Settings page. A modal dialog titled "Generate new token" is open. In the "Comment" field, the value "streetsmarttableaurspark" is entered. The "Lifetime (days)" dropdown is set to 90. The modal contains instructions and configuration values:

- Set new config values (leave input empty to accept default):
 - Databricks Host [no current value, must start with https://]: <https://adb-4002832894225232.12.azuredatabricks.net/?o=4002832894225232#setting/clusters/1105-063159-7j9yi5hs>
 - *** IMPORTANT: For AAD token users, please leave this empty and set AAD token via spark conf, spark.databricks.service.token
- Databricks Token [no current value]: dapi764addac1cfb0498c99c860714c4b9e0-3
- *** IMPORTANT: please ensure that your cluster has:
 - Databricks Runtime version of DBR 5.1+
 - Python version same as your local Python (i.e., 2.7 or 3.5)
 - the Spark conf `spark.databricks.service.enabled` set
- Cluster ID (e.g., 0921-001415-jelly628) [no current value]: 1105-063159-7j9yi5hs
- Org ID (Azure-only, see ?o=orgId in URL) [0]: 4002832894225232
- Port [15001]: 15001

...The Devil is in the Setup Details



...The Devil is in the Setup Details



The screenshot shows a RStudio interface with two main panes. The left pane is a 'Console' window displaying R code and its output. The right pane is a 'File Browser' showing a directory structure.

```
22  url <- dapply(df, function(x) x, schema))
23  collect(df1)
24
25
26
27
28:1 | (Top Level) ◆
Console Terminal × Background Jobs ×
R 4.2.1 · ~/...
> Sys.setenv(SPARK_HOME = "C:\Users\andre\anaconda3\envs\streetsmarttableaurspark\lib\site-packages\pyspark\jar")
Error: '\U' used without hex digits in character string starting "'C:\U"
> Sys.setenv(SPARK_HOME = "C:\Users\andre\anaconda3\envs\streetsmarttableaurspark\lib\site-packages\pyspark\jar")
Error: '\U' used without hex digits in character string starting "'C:\U"
> Sys.setenv(SPARK_HOME = "C:\Users\andre\anaconda3\envs\streetsmarttableaurspark\lib\site-packages\pyspark\jar")
Error: malformed raw string literal at line 1
> Sys.setenv(SPARK_HOME = "C:\Users\andre\anaconda3\envs\streetsmarttableaurspark\lib\site-packages\pyspark\jar")
> sparkR.session()
Spark not found in SPARK_HOME: C:\Users\andre\anaconda3\envs\streetsmarttableaurspark\lib\site-packages\pyspark\jar
Will you download and install (or reuse if it exists) Spark package under the cache [C:\Users\andre\AppData\Local\Apache\Spark\Cache]? (y/n): y
Will you download and install (or reuse if it exists) Spark package under the cache [C:\Users\andre\AppData\Local\Apache\Spark\Cache]? (y/n): d
f <- as.DataFrame(faithful)
Will you download and install (or reuse if it exists) Spark package under the cache [C:\Users\andre\AppData\Local\Apache\Spark\Cache]? (y/n): h
head(df)
Will you download and install (or reuse if it exists) Spark package under the cache [C:\Users\andre\AppData\Local\Apache\Spark\Cache]? (y/n): h
Will you download and install (or reuse if it exists) Spark package under the cache [C:\Users\andre\AppData\Local\Apache\Spark\Cache]? (y/n): d
f1 <- dapply(df, function(x) x, schema(df))
Will you download and install (or reuse if it exists) Spark package under the cache [C:\Users\andre\AppData\Local\Apache\Spark\Cache]? (y/n): c
collect(df1)
Will you download and install (or reuse if it exists) Spark package under the cache [C:\Users\andre\AppData\Local\Apache\Spark\Cache]? (y/n): s
ys.setenv(SPARK_HOME = "C:\Users\andre\anaconda3\envs\streetsmarttableaurspark\lib\site-packages\pyspark")
Will you download and install (or reuse if it exists) Spark package under the cache [C:\Users\andre\AppData\Local\Apache\Spark\Cache]? (y/n): s
sparkR.session()
Will you download and install (or reuse if it exists) Spark package under the cache [C:\Users\andre\AppData\Local\Apache\Spark\Cache]? (y/n): y
Spark not found in the cache directory. Installation will start.
MirrorUrl not provided.
Looking for preferred site from apache website...
Preferred mirror site found: https://dlcdn.apache.org/spark
Downloading spark-3.2.2 for Hadoop 2.7 from:
- https://dlcdn.apache.org/spark/spark-3.2.2/spark-3.2.2-bin-hadoop2.7.tgz
trying URL 'https://dlcdn.apache.org/spark/spark-3.2.2/spark-3.2.2-bin-hadoop2.7.tgz'
Content type 'application/x-gzip' length 272846416 bytes (260.2 MB)
```

Download progress
https://dlcdn.apache.org/spark/spark-3.2.2/spark-3.2.2-bin-hadoop2.7.tgz
e=0, modificationTime=16511
ESS: The
of PID
of PID
indows\System32>
indows\System32>dataricks-
streetsmarttableaurspark\lib
indows\System32>
like these other systems, RCloud:
• lets you easily browse and search
them, and use them as function c
• lets you interpret notebooks as w
automated dashboard.
• provides an environment in which
• provides a transparent, integrated
you need low-level access to RCloud
because RCloud notebooks are G
erested? Try RCloud on the public i
GitHub, Inc. Terms Privacy

...The Devil is in the Setup Details



```
(streetSMARTTableauSpark) C:\Windows\System32>databricks-connect test
* PySpark is installed at C:\Users\andre\anaconda3\envs\streetSMARTTableauSpark\lib\site-packages\pyspark
* Checking SPARK_HOME
* Checking java version
openjdk version "17.0.3" 2022-04-19 LTS
OpenJDK Runtime Environment Zulu17.34+19-CA (build 17.0.3+7-LTS)
OpenJDK 64-Bit Server VM Zulu17.34+19-CA (build 17.0.3+7-LTS, mixed mode, sharing)
WARNING: Java versions >8 are not supported by this SDK
* Skipping scala command test on Windows
* Testing python command
22/11/05 02:01:33 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
Using Spark's default log4j profile: org/apache/spark/log4j-defaults.properties
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).
22/11/05 02:01:33 WARN MetricsSystem: Using default name SparkStatusTracker for source because neither spark.metrics.namespace nor spark.app.id is set.
View job details at https://adb-4002832894225232.12.azure.databricks.net/?o=4002832894225232&o=4002832894225232#/settings/clusters/1105-063159-7j9yi5hs/sparkUi
* Simple PySpark test passed
* Testing dbutils.fs
[FileInfo(path='dbfs:/FileStore/', name='FileStore/', size=0, modificationTime=1654628862000), FileInfo(path='dbfs:/databricks-datasets/', name='databricks-datasets/', size=0, modificationTime=0), FileInfo(path='dbfs:/databricks-results/', name='databricks-results/', size=0, modificationTime=0), FileInfo(path='dbfs:/delta/', name='delta/', size=0, modificationTime=1653678776000), FileInfo(path='dbfs:/local_disk0/', name='local_disk0/', size=0, modificationTime=1665054121000), FileInfo(path='dbfs:/mnt/', name='mnt/', size=0, modificationTime=1651155365000), FileInfo(path='dbfs:/outputs/', name='outputs/', size=0, modificationTime=1654628727000), FileInfo(path='dbfs:/scripts/', name='scripts/', size=0, modificationTime=1654629190000), FileInfo(path='dbfs:/tmp/', name='tmp/', size=0, modificationTime=1664423920000), FileInfo(path='dbfs:/user/', name='user/', size=0, modificationTime=1651155363000)]
* Simple dbutils test passed
* All tests passed.
```

Demo:
Local R & Remote R Integration





Street Smart Machine Learning



Machine Learning Process

Model Production

Data Gathering

Feature Engineering

Sampling and Splitting

Modeling

Validating

Scoring

Drift Management

Monitor Live Scoring

Pipeline Development

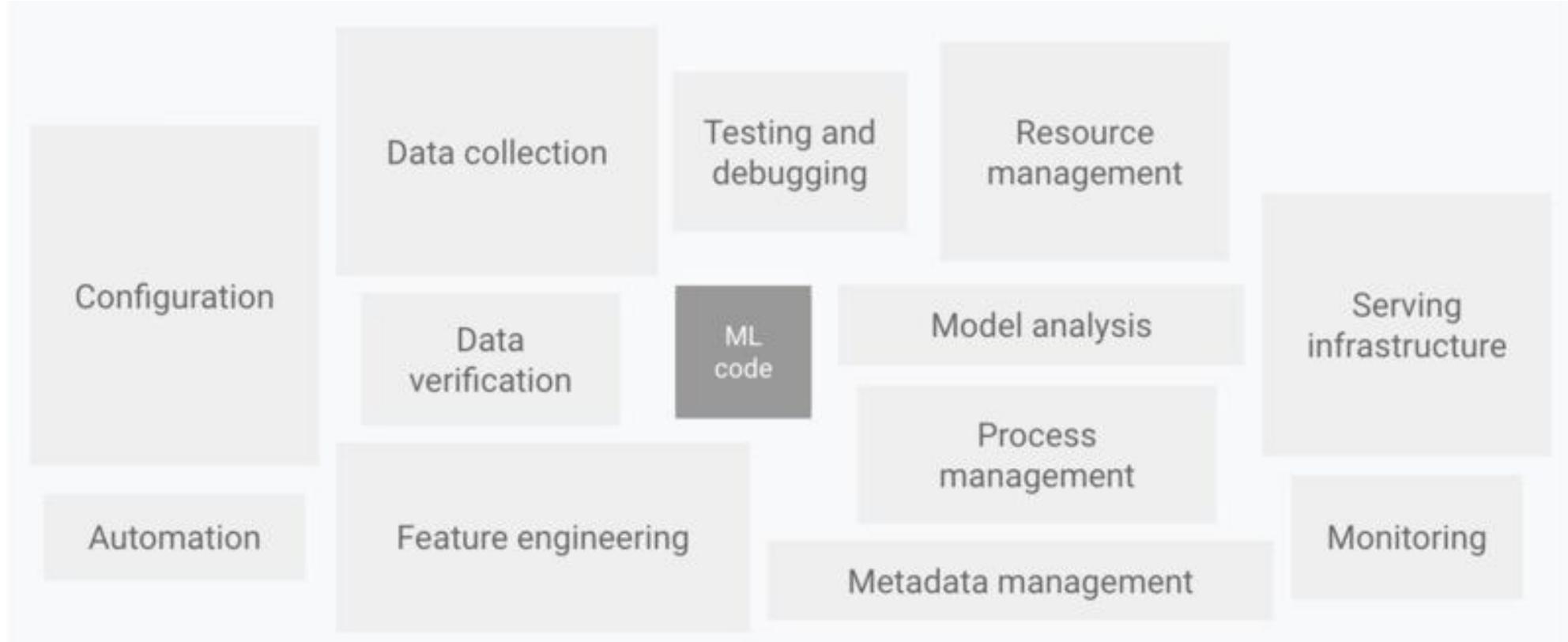
Production Gap Analysis

Business Alignment

Model Consumption

Advanced Analytics – Technical Debt

A Proportional Measure of Estimated Effort in the Entire Machine Learning Lifecycle



Machine Learning Modeling Families



Estimation



Association



Clustering



Classification Forecasting



Machine Learning Modeling Families



Estimation



Association



Clustering



Classification Forecasting



Target a Discrete Answer –Yes/No

- Find All Columns Driving its Value
- Use model to score new records

Machine Learning Modeling Families



Estimation



Association



Clustering



Classification



Forecasting

- Hard and Soft Groupings
- Profiles of Subgroups
- Likenesses and Differences

Machine Learning Modeling Families



Estimation



Association



Clustering

Predicting a Continuous Distribution
▪ Many Different Measures of Accuracy



Classification Forecasting



Machine Learning Modeling Families



Estimation



Association

- Collaborative Filtering
- Identify cross-sell
- Identify sequential, next-sale
- Make purchase recommendations
- Complex event associations



Clustering



Classification



Forecasting

Machine Learning Modeling Families



Estimation



Association



Clustering



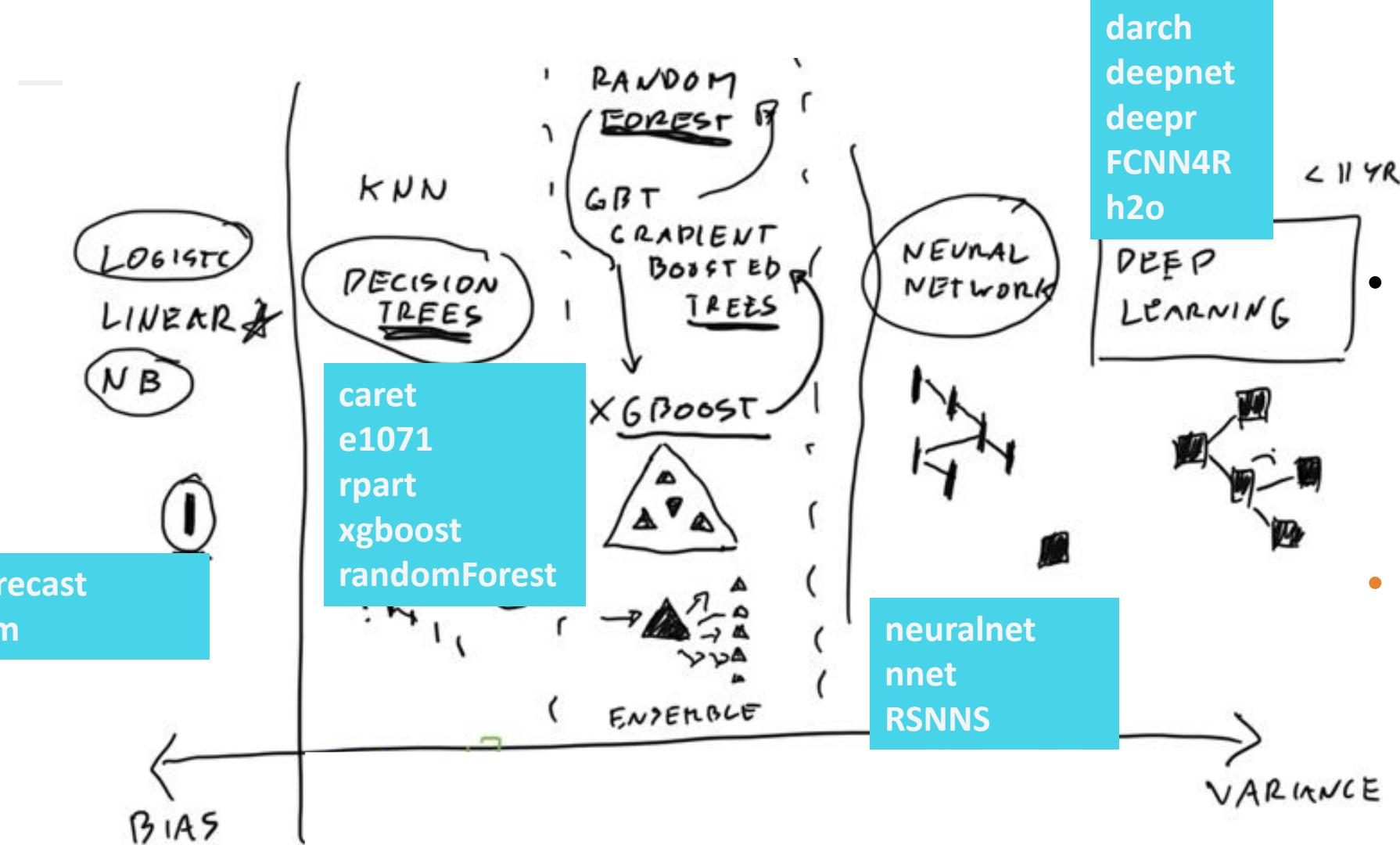
Classification



Forecasting

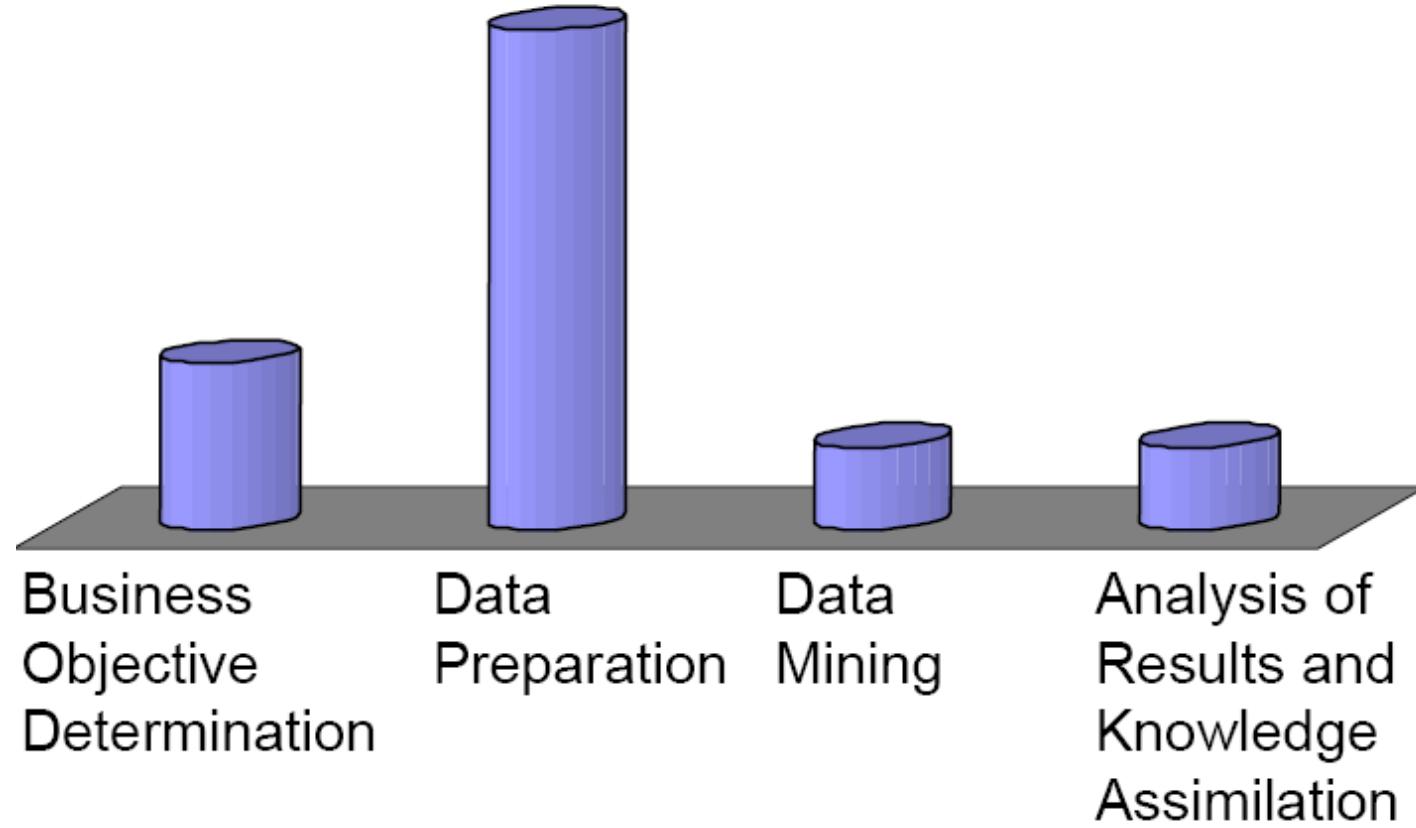
- Input of measure over time and related series
- Predictions generated for short term trends
- Based on cycles and events

Machine Learning Algorithm Complexity & Context



- LR, NN and other algorithms part of **hierarchical layers** of DL model
- **Transfer Learning**
Learning models on same type of data

Advanced Analytics Task Breakout



Experimental Design – On Paper



THE DATA SCIENCE **HIERARCHY OF NEEDS**

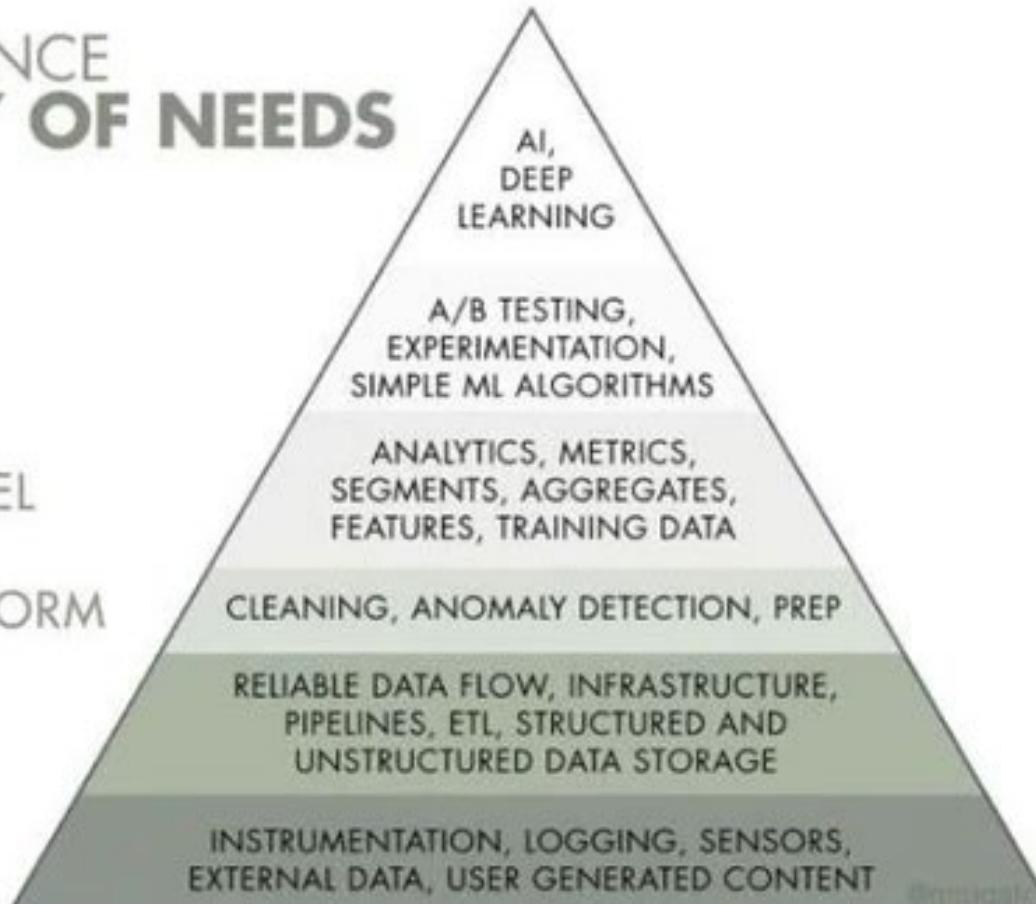
LEARN/OPTIMIZE

AGGREGATE/LABEL

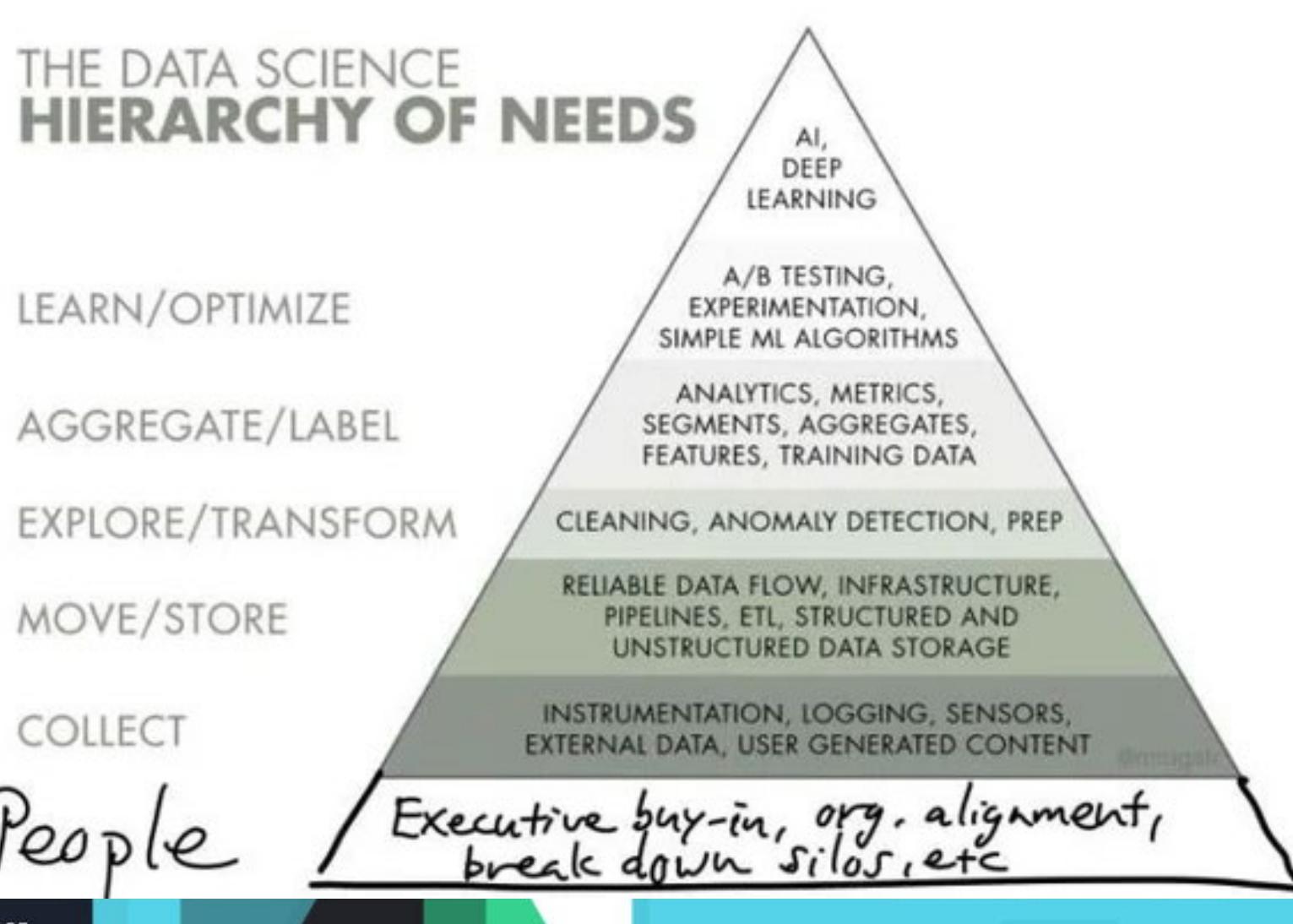
EXPLORE/TRANSFORM

MOVE/STORE

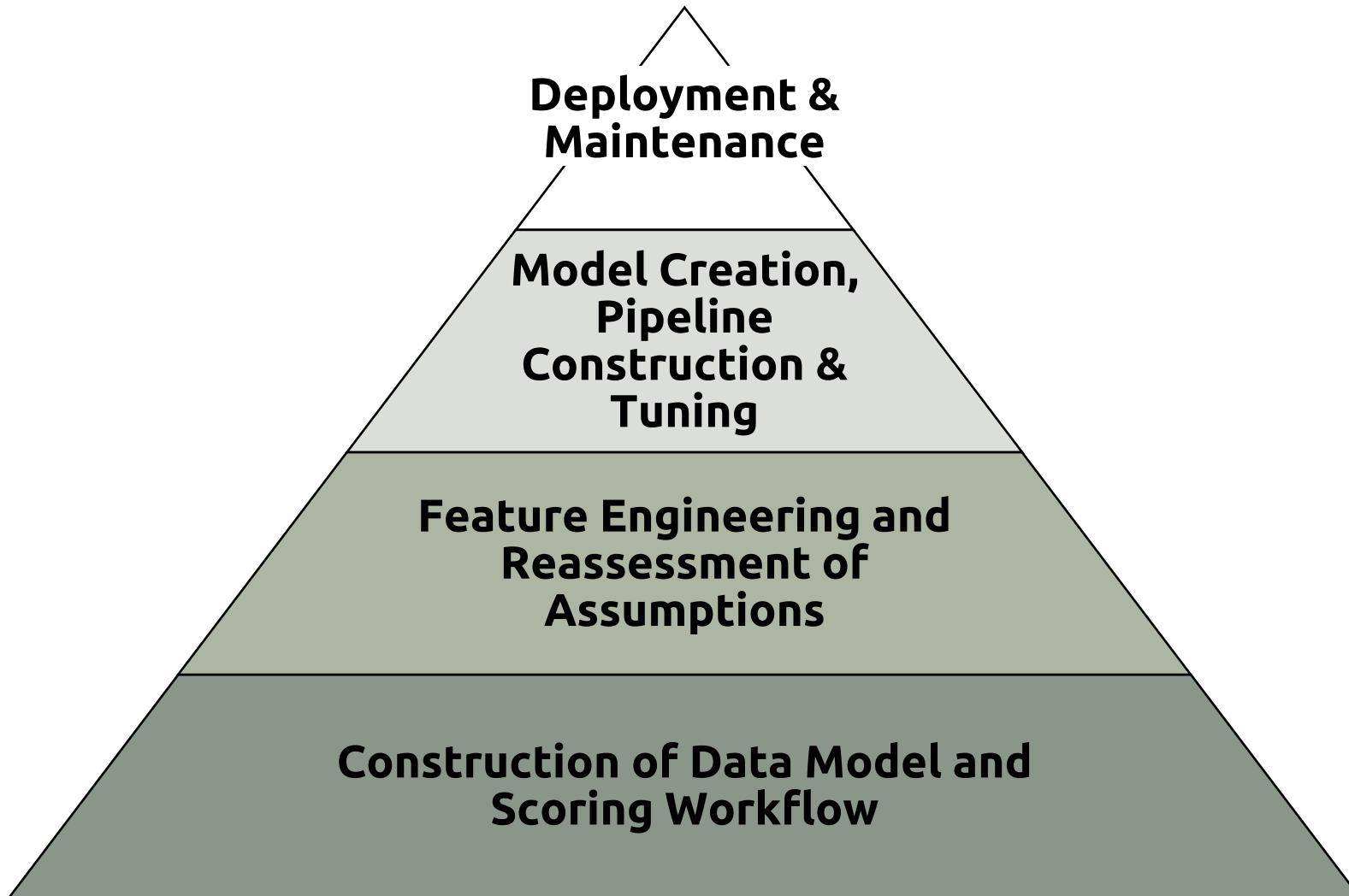
COLLECT



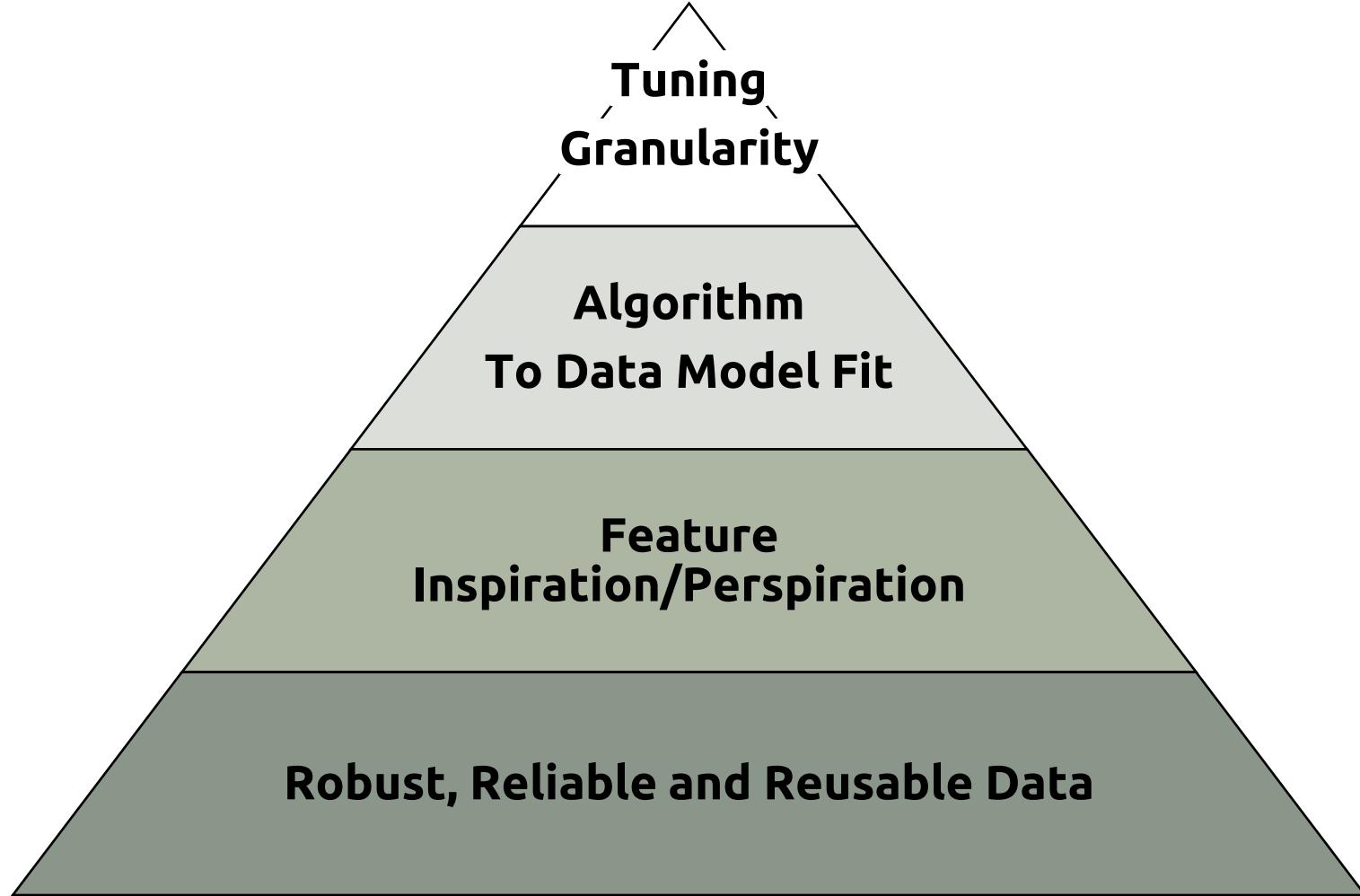
Experimental Design – In Real Life



Experimental Design – Effort By % of Labor



Experimental Design – Effort ROI





Thank you!!!

Deck & Repo at

<https://github.com/hekaplex/StreetSmartTableauRSpark>

DIVERGENCE ACADEMY

(214) 766-3897

drew.minkin@divergence.one

