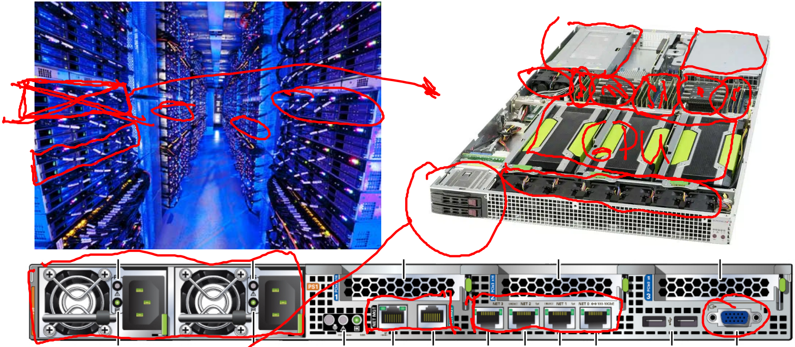
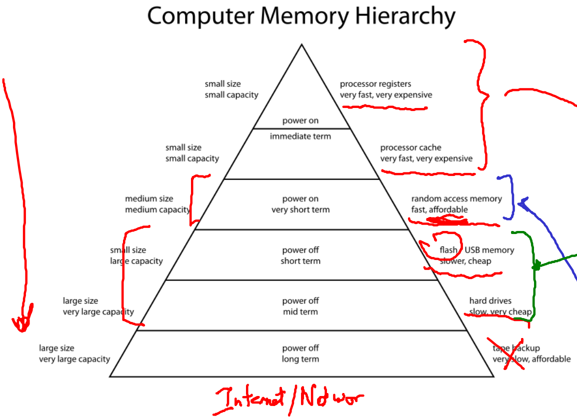
A Primer on Computers and Data Centers

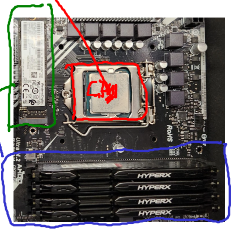


- Dual socket CPU: 96+ cores each
- Memory: 128GB to 1TB
- Storage: 500GB SSD plus multiple SSDs/HDDs for data storage

This is only about 5-20 times as powerful as your computer . . .

The Computer Memory Hierarchy



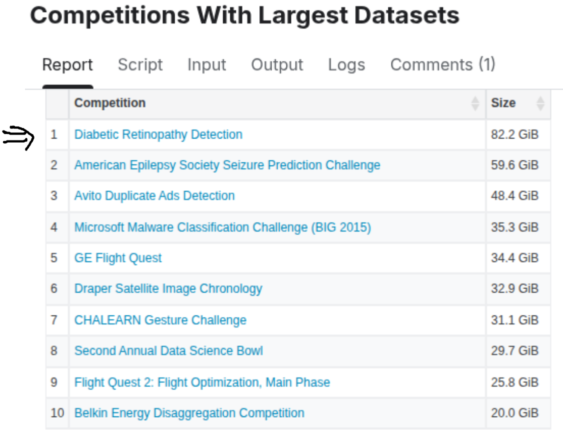


Key point:

- To be very fast, the 'inner loop' must fit inside cache
- To be fast enough, the 'inner loop' must fit inside RAM

Big Data

Problem: There are some really big datasets out there . . .



Competitions With Largest Datasets

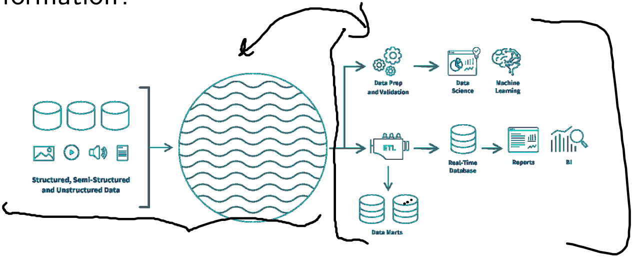
	Competition	Size
1	Diabetic Retinopathy Detection	82.2 GiB
2	American Epilepsy Society Seizure Prediction Challenge	59.6 GiB
3	Avito Duplicate Ads Detection	48.4 GiB
4	Microsoft Malware Classification Challenge (BIG 2015)	35.3 GiB
5	GE Flight Quest	34.4 GiB
6	Draper Satellite Image Chronology	32.9 GiB
7	CHALEARN Gesture Challenge	31.1 GiB
8	Second Annual Data Science Bowl	29.7 GiB
9	Flight Quest 2: Flight Optimization, Main Phase	25.8 GiB
10	Belkin Energy Disaggregation Competition	20.0 GiB

Another page claims there are 526 datasets on Kaggle $>$ 100GB

Data Analytics with Spark

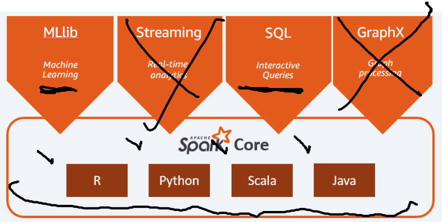
"Data analytics converts raw data into actionable insights."

Problem: How to convert an organization's huge data sets into useful information?



- "Apache Spark is an open-source, distributed processing system used for big data workloads."
- "It utilizes in-memory caching, and optimized query execution for fast analytic queries against data of any size."
- "Spark provides an interface for programming clusters with implicit data parallelism and fault tolerance."

Apache Spark Internals



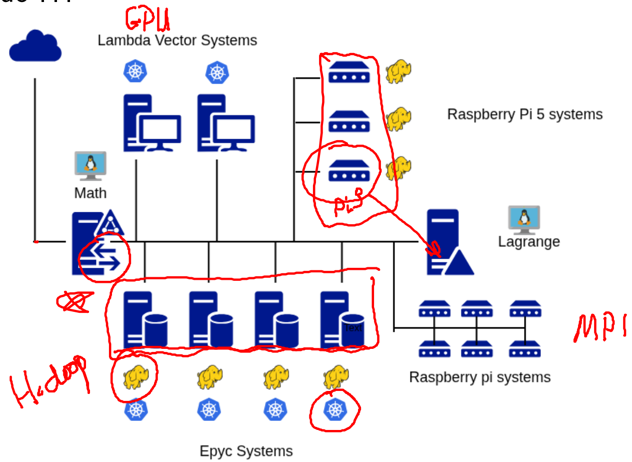
You

- Spark is built on Hadoop Mapreduce, “a programming model …for processing and generating big data sets with a parallel and distributed algorithm on a cluster.”
- Spark “has its architectural foundation in …a read-only multiset of data items distributed over a cluster of machines, that is maintained in a fault-tolerant way.” [Dataframes]
- Spark uses HDFS, which “provides the scalable, fault-tolerant storage layer, while Spark acts as the high-speed, in-memory data processing engine that operates on the data stored in HDFS.”

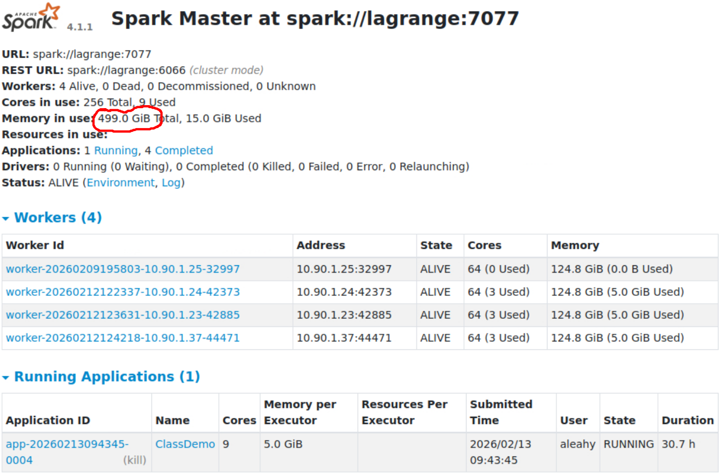
Apache Spark at Knox

Question: Why have you never used Apache Spark? Because you don't have a cluster . . .

but we do . . .



Apache Spark at Knox



Spark Master at spark://lagrange:7077

4.1.1

URL: spark://lagrange:7077
REST URL: spark://lagrange:6066 *(cluster mode)*
Workers: 4 Alive, 0 Dead, 0 Decommissioned, 0 Unknown
Cores in use: 256 Total, 9 Used
Memory in use: 499.0 GiB Total, 15.0 GiB Used
Resources in use:
Applications: 1 Running, 4 Completed
Drivers: 0 Running (0 Waiting), 0 Completed (0 Killed, 0 Failed, 0 Error, 0 Relaunching)
Status: ALIVE (Environment, Log)

Workers (4)

Worker Id	Address	State	Cores	Memory
worker-20260209195803-10.90.1.25-32997	10.90.1.25:32997	ALIVE	64 (0 Used)	124.8 GiB (0.0 B Used)
worker-20260212122337-10.90.1.24-42373	10.90.1.24:42373	ALIVE	64 (3 Used)	124.8 GiB (5.0 GiB Used)
worker-20260212123631-10.90.1.23-42885	10.90.1.23:42885	ALIVE	64 (3 Used)	124.8 GiB (5.0 GiB Used)
worker-20260212124218-10.90.1.37-44471	10.90.1.37:44471	ALIVE	64 (3 Used)	124.8 GiB (5.0 GiB Used)

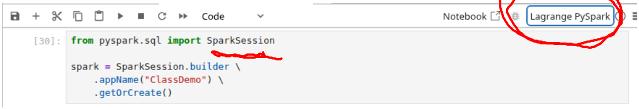
Running Applications (1)

Application ID	Name	Cores	Memory per Executor	Resources Per Executor	Submitted Time	User	State	Duration
app-20260213094345-0004 (kill)	ClassDemo	9	5.0 GiB		2026/02/13 09:43:45	aleahy	RUNNING	30.7 h

This cluster uses an 8TB HDFS filesystem with double redundancy on each file block spread over the same four systems

Using Spark: Basic Concepts

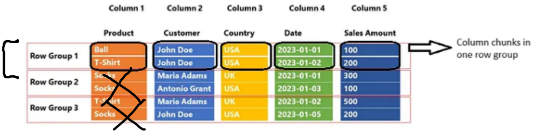
- A Spark cluster can be accessed with a *SparkSession* from several programming languages (Python, R, Scala, . . .)



```
[30]: from pyspark.sql import SparkSession

      spark = SparkSession.builder \
          .appName("ClassDemo") \
          .getOrCreate()
```

- Spark typically uses the *parquet* file format



	Column 1	Column 2	Column 3	Column 4	Column 5
	Product	Customer	Country	Date	Sales Amount
Row Group 1	Ball	John Doe	USA	2023-01-01	100
	T-Shirt	John Doe	USA	2023-01-02	200
Row Group 2	Socks	Maria Adams	UK	2023-01-01	300
	Socks	Antonio Grant	USA	2023-01-03	100
Row Group 3	T-Shirt	Maria Adams	UK	2023-01-02	500
	Socks	John Doe	USA	2023-01-05	200

ID, sex, number,

  - Parquet is a *columnar* file format, which allows *pruning*
  - Columnar data can be compressed, because it is similar
  - Parquet supports *predicate pushdown*, in which a *where* filter "filters the data in the database query"
  - Spark supports all this, which speeds up queries up to 10-100x

The Spark API

Key point: Spark has to solve the same problems as Pandas does . . . See the Spark API reference (for 4.1.0)

Another key point: The Spark API has been evolving, so there are multiple entrypoints–some better than others

Some Entrypoints for the API:

- Spark SQL - data filtering and aggregating
- Pandas API - data input/output and transformation
- MLlib - many familiar machine learning models

See the Spark API Reference:

https://spark.apache.org/docs/latest/api/python/reference/index.html

. . . or try your favorite AI for example syntax . . .