“Transform **Real** Time Data into **Real** Time Decisions”

Asit Parija(@asitparija)

SIGMOID ANALYTICS

OPEN SOURCE

CUSTOMERS

PubNub

guavus

CAPILLARY™

fusionops

PARTNERSHIPS

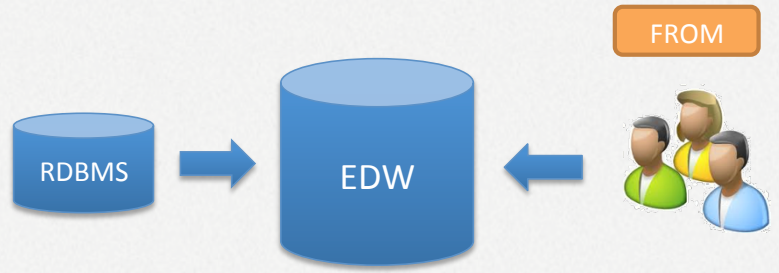
cloudera

DATABRICKS

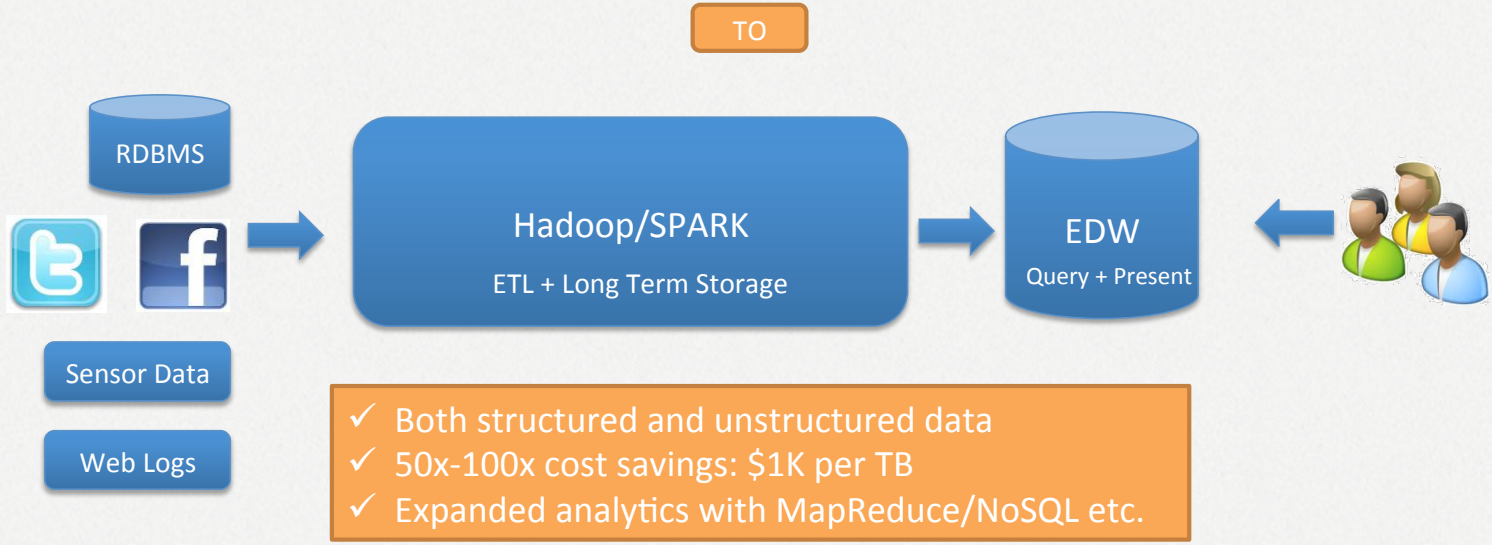
MAPR™ TECHNOLOGIES

ITC INFOTECH
Business-friendly Solutions

ETL



- Only structured data
- $50K – 100K per TB
- Limited Analytics



- ✓ Both structured and unstructured data
- ✓ 50x-100x cost savings: $1K per TB
- ✓ Expanded analytics with MapReduce/NoSQL etc.

ETL Goals

SIGMOID ANALYTICS

- Make data processing more powerful
- Make data processing more simple
- Make data processing 100x faster than before
- What are the options ?

What steered us into Spark

- Powerful in-memory Processing
- Simple operator on Data
- Debuggable API
- Efficient Execution
- Universally distributed

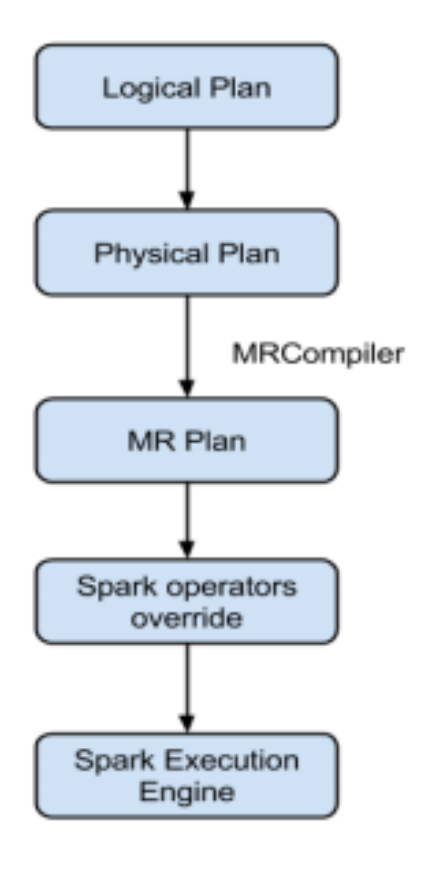
What steered us into Pig

- DSL for ETL
- Rich Operator Library
- Extendable
- Pluggable
- Powerful ETL

Operator Mapping

Pig	Spark
Load	HadoopRDD
Store	saveasObjectFile
Filter	MappedRDD + filter func
GroupBY (Local rearrange, global rearrange & package)	Sort + Group by
….	…

Current Flow



Issues

- Scaling
- Performance
- Spark Specific Operators (Cache)
- Pig on Spark Unit test
- Some specific joins & rank operation

Filter Code implementation

- https://bitbucket.org/SigmoidDev/spork/src/80a3e4626e4504c1829568942e0690abc79d239a/src/org/apache/pig/backend/hadoop/executionengine/spark/converter/FilterConverter.java?at=spork-1.0

Contribute

- Pig on Spark Umbrella Jira
- https://issues.apache.org/jira/browse/PIG-4059

- https://github.com/sigmoidanalytics/spork
- Issues

Benchmark

Distinct operation on the data is a wikistats dump for 25 days with size 270G took 4.25mins on Pig on Spark, as compared to 30mins in MapReduce .

Mixing Streaming & Batch Processing

- Current State – Different code for batch and stream
- Lambda Architecture
- One unified language to perform both

What else is cool

CloudFlux	**SigmaStream**
Cloud Deployment	PIG/SQL Like DSL
Fault Tolerance	Rich Stream operators
AutoScaling	Multiple Data source/Sink
Programmatic interface	Add custom Operators
Cloud Agnostic	Apache Spark Based
Apache License	Apache License

Thank You

US Office

1343 Kingfisher Way
Sunnyvale, CA, 94087

+1 (760) 203 3257

contact@sigmoidanalytics.com

India Office

Gulmohar Enclave Road,
Silver Spring Layout, Munnekollal
Bengaluru, Karnataka 560037