DataTorrent

Unified Batch & Stream Processing Platform

Himanshu Bari

Director Product Management

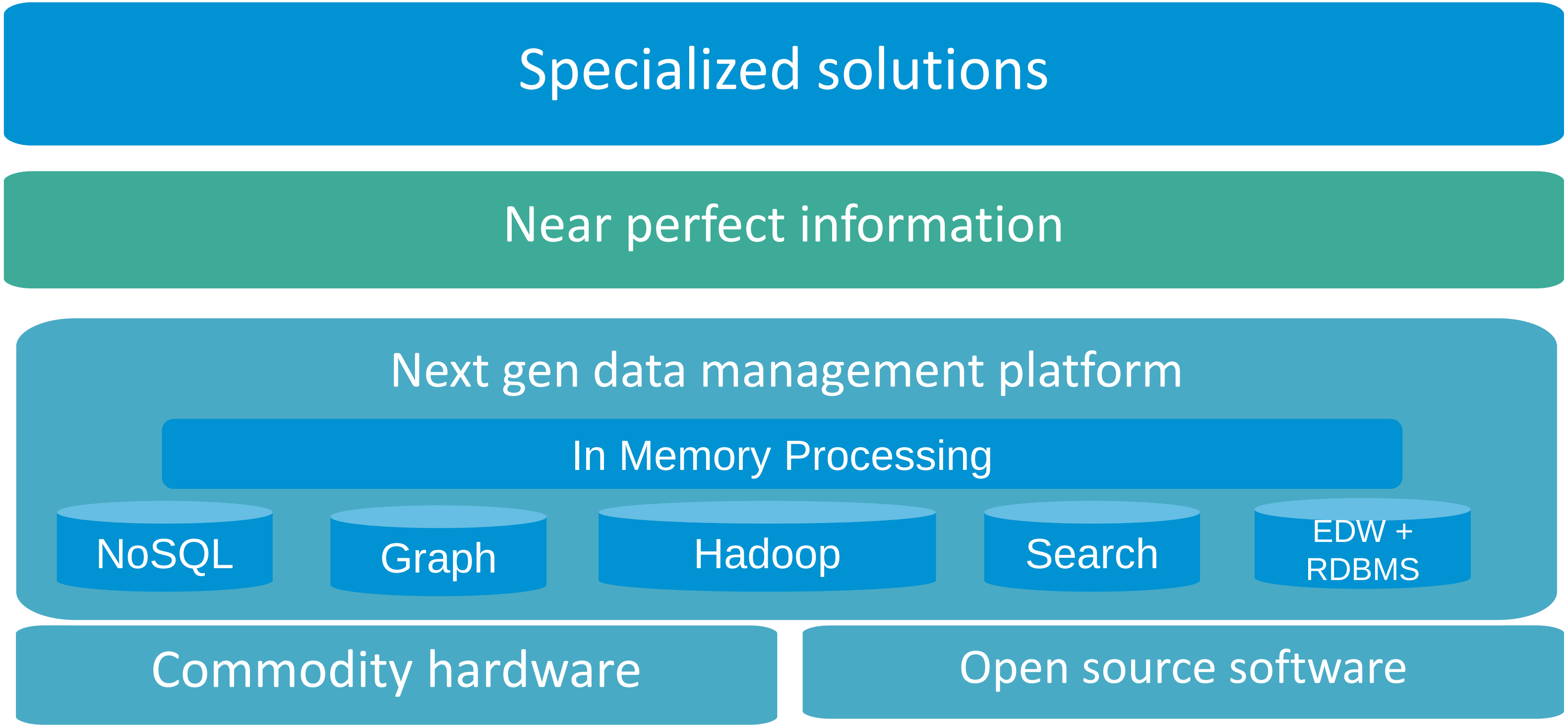
Most Big Data Use Cases Are About

Improving/Re-write EXISTING solutions To KNOWN problems…

Current Solutions Were Built On

A. Imperfect information
B. Expensive s/w & h/w infrastructure
C. Relational data stores

Inevitable Course of the Re-write



Specialized solutions

Near perfect information

Next gen data management platform

In Memory Processing

NoSQL

Graph

Hadoop

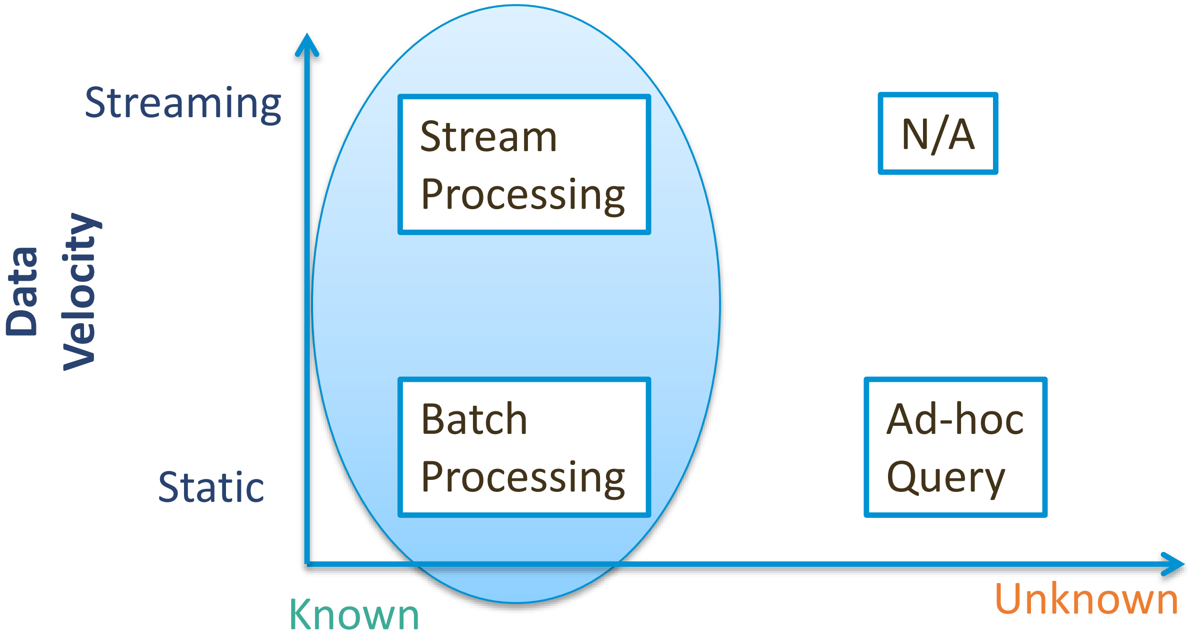
Search

EDW + RDBMS

Commodity hardware

Open source software

Data Processing Categories in Big Data Use Cases



Questions known?

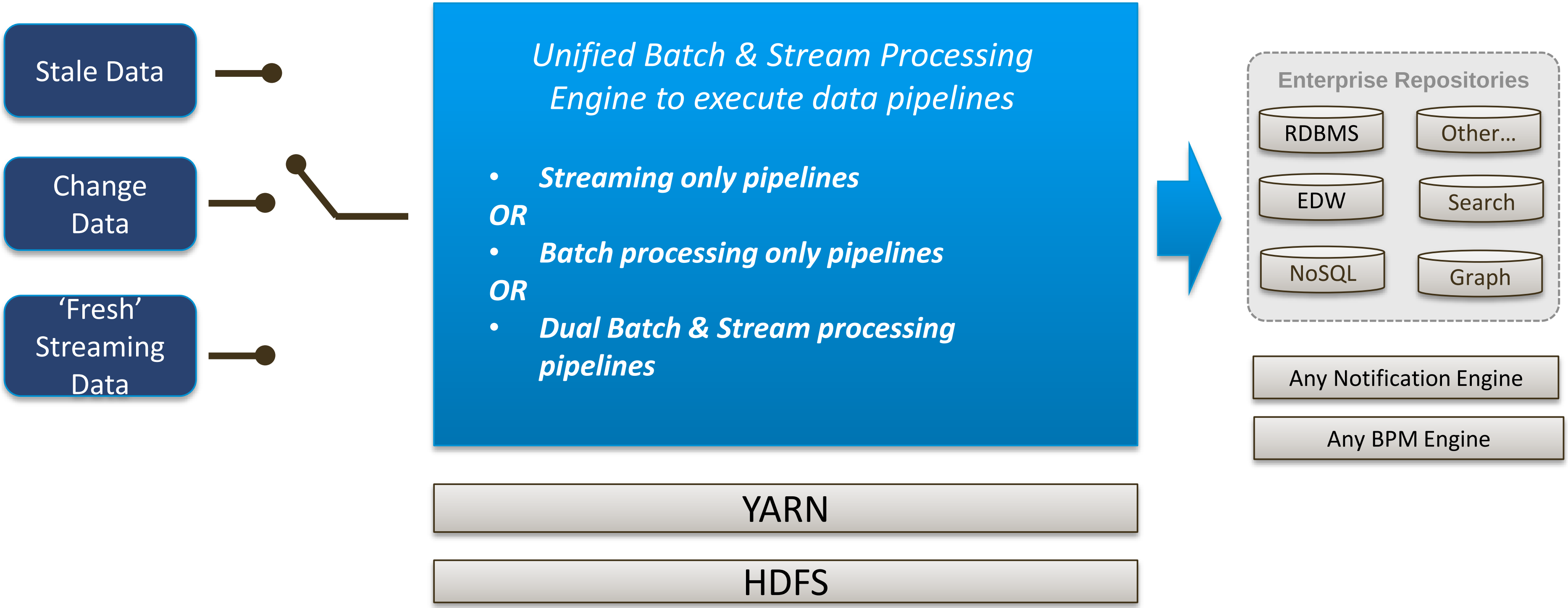
Every Batch Process *Could* Have Been A Stream Process

- Every ‘Static’ data point was ‘Streaming’ at some point
- We choose to wait and collect a bunch of data points and then process them at once in ‘Batch Mode’
- Move processing time closer to the data generation or ‘Event’ occurrence time
- Reduces time to insight and allows you to be proactive with timely actions

But We Still Need Batch

- Historical data analysis
- What-if analysis
- Experimentation
- Data re-statement
- Transaction processing and re-conciliation
- Audit
- Machine learning model training
- ……..

Need A Unified Platform



Stale Data

Change Data

'Fresh' Streaming Data

Unified Batch & Stream Processing Engine to execute data pipelines

- ***Streaming only pipelines***

OR

- ***Batch processing only pipelines***

OR

- ***Dual Batch & Stream processing pipelines***

Enterprise Repositories

RDBMS

Other…

EDW

Search

NoSQL

Graph

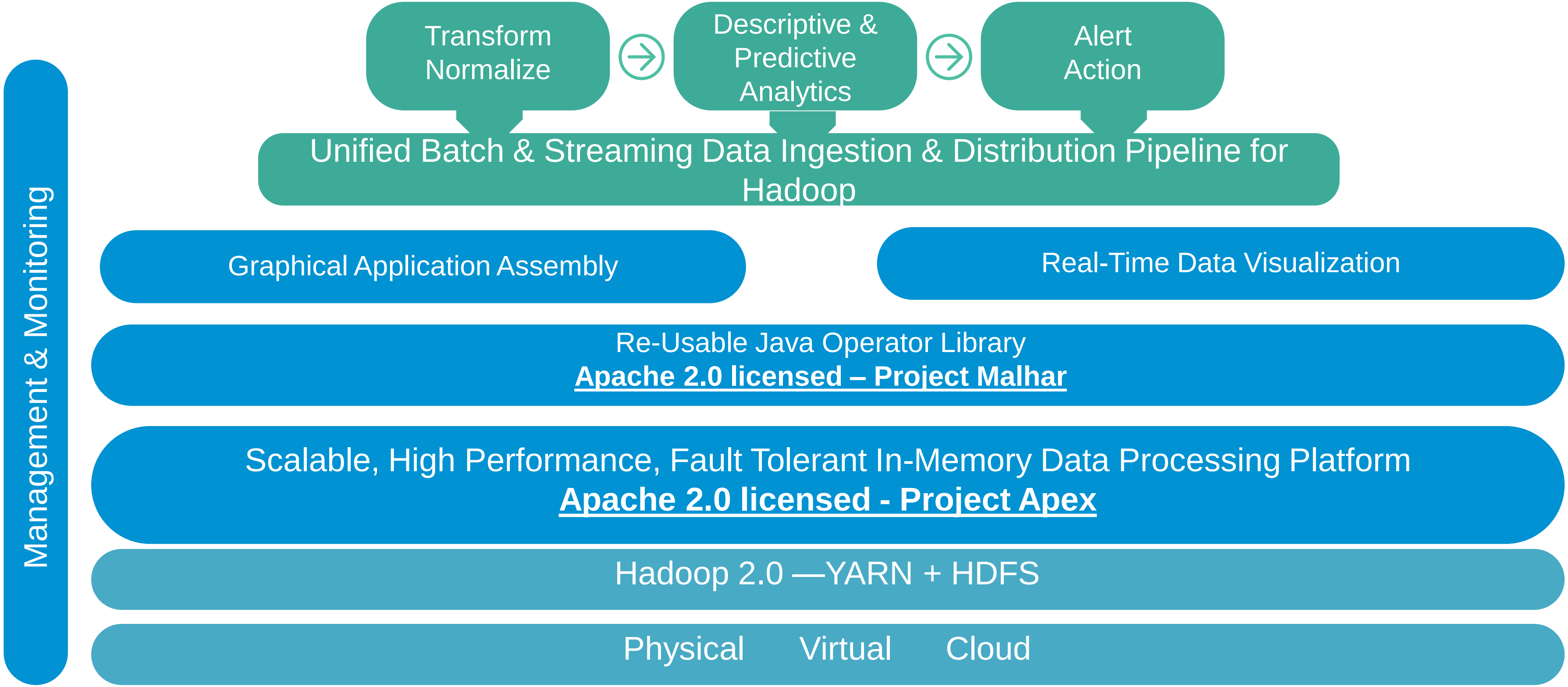
Any Notification Engine

Any BPM Engine

YARN

HDFS

DataTorrent RTS Provides A Unified Batch & Streaming Platform



Transform Normalize

Descriptive & Predictive Analytics

Alert Action

Unified Batch & Streaming Data Ingestion & Distribution Pipeline for Hadoop

Graphical Application Assembly

Real-Time Data Visualization

Re-Usable Java Operator Library
Apache 2.0 licensed – Project Malhar

Scalable, High Performance, Fault Tolerant In-Memory Data Processing Platform
Apache 2.0 licensed - Project Apex

Hadoop 2.0 —YARN + HDFS

Physical Virtual Cloud

Management & Monitoring

DataTorrent Project Apex- Unified Batch & Stream Processing Engine

1 In-memory compute engine

2 Hadoop native (YARN + HDFS) architecture

3 Sub-second event processing

4 Event order guarantees

5 Flexible stream partitioning mechanism

6 Stateful fault tolerance with NO data loss

7 Automatic & incremental recovery

8 Rolling & tumbling windows

9 Processing guarantees

10 Ability to update application without downtime

Unified Data Ingestion & Distribution Pipeline

Input & Output Variety

- FTP, S3 etc.
- Kafka & JMS
- Change data capture

Tackle data size & speed fluctuations

- Overcome HDFS small file problem
- Skew management

Hadoop Native

- Runs within the Hadoop cluster over YARN & HDFS

Easily customizable

- Easily extend and insert operations for data preparation

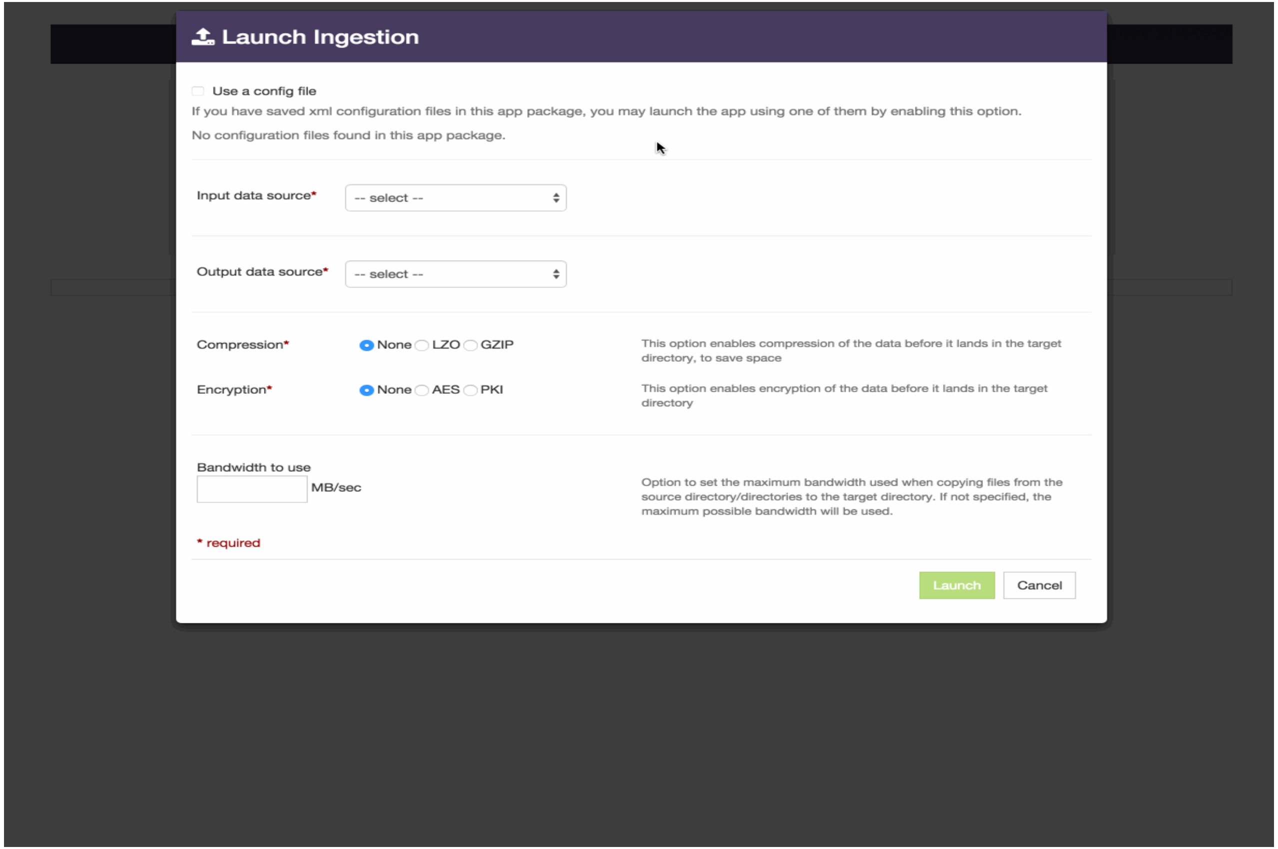
Run-time updates

- Parameters like filtering criteria, bandwidth utilization & polling interval should be updateable at runtime

Simple to build & Operate

- Graphical UI & API
- End to end metrics

Hadoop Data Ingestion & Distribution Application



Launch Ingestion

□ Use a config file

If you have saved xml configuration files in this app package, you may launch the app using one of them by enabling this option.

No configuration files found in this app package.

Input data source* -- select --

Output data source* -- select --

Compression* ● None ○ LZO ○ GZIP

This option enables compression of the data before it lands in the target directory, to save space

Encryption* ● None ○ AES ○ PKI

This option enables encryption of the data before it lands in the target directory

Bandwidth to use MB/sec

Option to set the maximum bandwidth used when copying files from the source directory/directories to the target directory. If not specified, the maximum possible bandwidth will be used.

* required

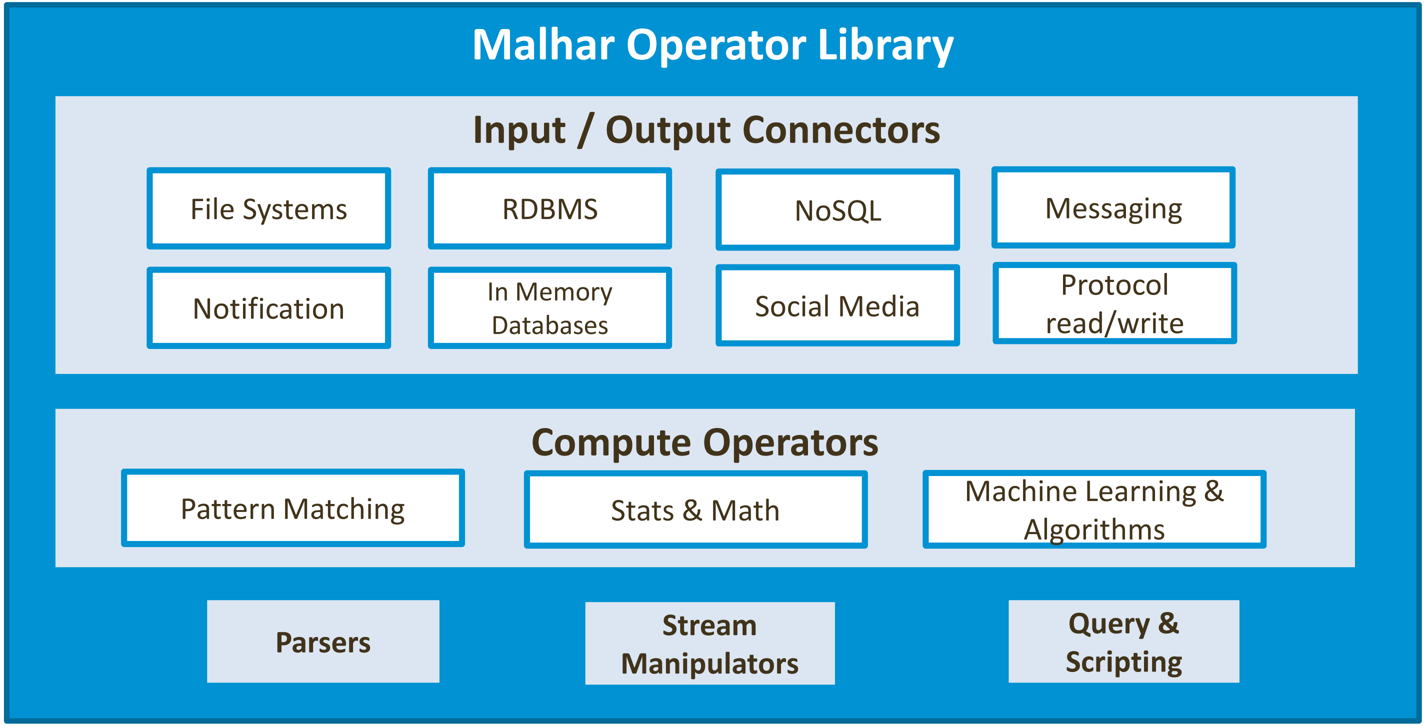
Data Prep & Analytics Layer Requirements

- Truly scale horizontally across the Hadoop cluster
- Pre-built operators
  - Re-ordering
  - Normalization
  - Transpose
  - De-duplication
  - Tagging
  - Filtering
  - Enrichment
- Operators work seamlessly in both streaming & batch mode
  - Data local HDFS read & process
  - Ability to do computations on time window as well as file boundaries
- Ability to re-use existing business logic
- Simple workflow & scheduling capabilities
  - Built-in or integrations with Oozie or other schedulers

Development API Requirements

- Consistent between Streaming & Batch pipelines
- No mapreduce
- No exposure of low level processing engine concepts
- Easily extendible

Malhar Operator Library Overview



Malhar Operator Library

Input / Output Connectors

- File Systems
- RDBMS
- NoSQL
- Messaging
- Notification
- In Memory Databases
- Social Media
- Protocol read/write

Compute Operators

- Pattern Matching
- Stats & Math
- Machine Learning & Algorithms

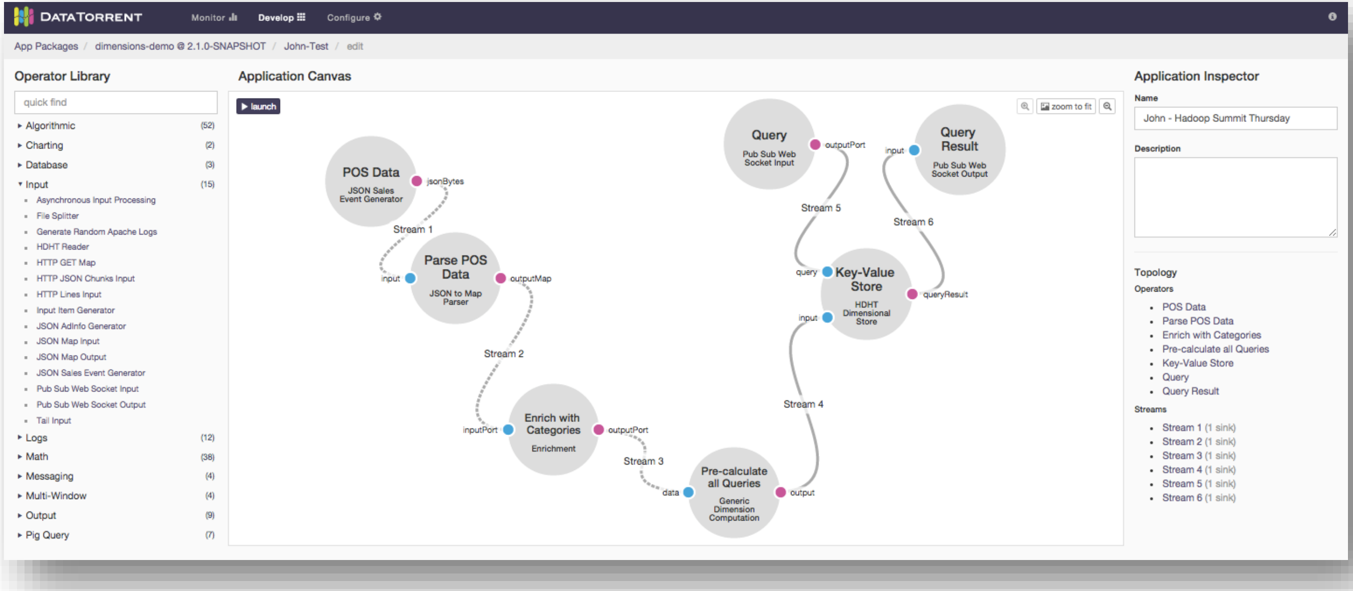
Parsers

Stream Manipulators

Query & Scripting

Visual Application Assembly

- Easy to Use
  - Web based drag-n-drop development environment
  - Automatic port compatibility validation
  - Simple schema management
  - Generic property configurator



Name
John - Hadoop Summit Thursday

Description

Topology

Operators

- POS Data
- Parse POS Data
- Enrich with Categories
- Pre-calculate all Queries
- Key-Value Store
- Query
- Query Result

Streams

- Stream 1 (1 sink)
- Stream 2 (1 sink)
- Stream 3 (1 sink)
- Stream 4 (1 sink)
- Stream 5 (1 sink)
- Stream 6 (1 sink)

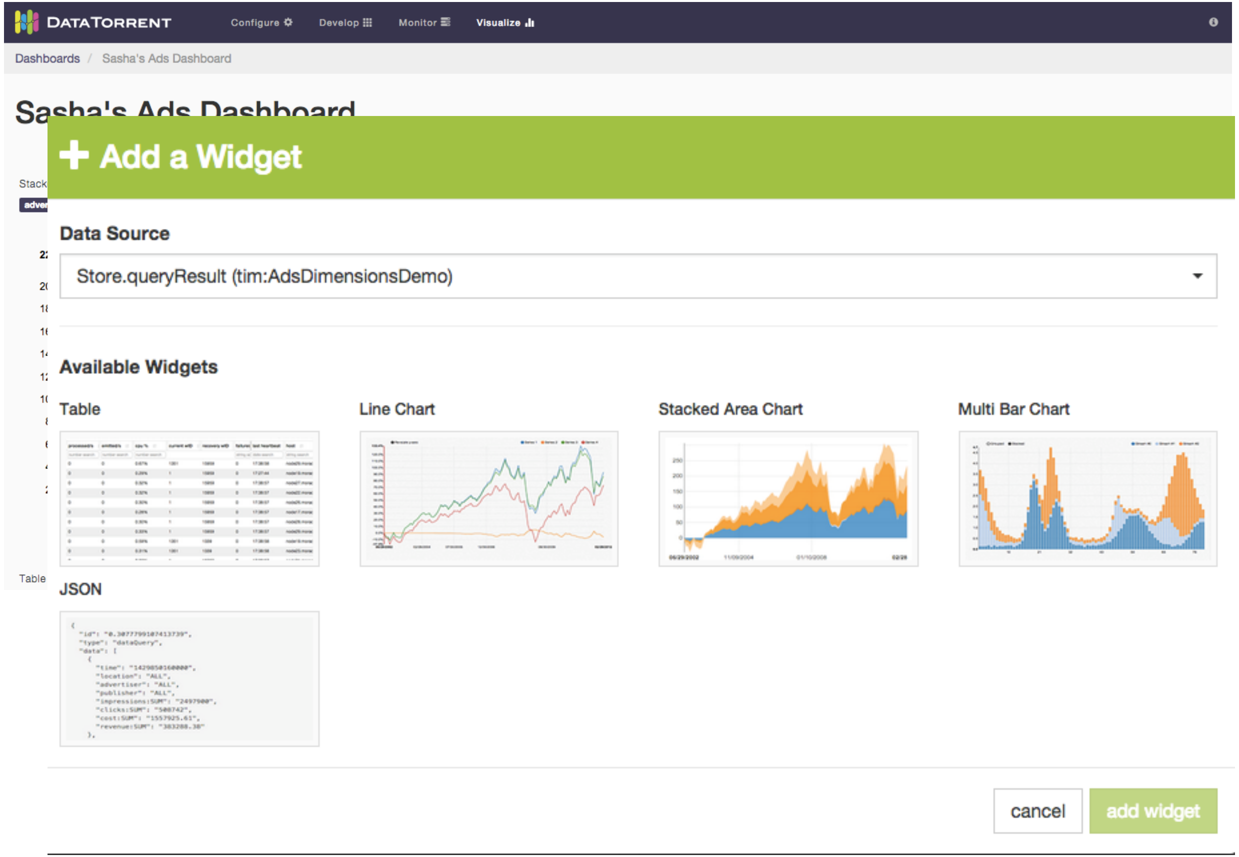
- Easy to Operate
  - No external component dependency - Runs natively in Hadoop
  - Integrated with DataTorrent management platform

- Simple to extend
  - Simple API to enable any existing DataTorrent operator
  - Ability to plug any business logic using a custom operator

DATATORRENT

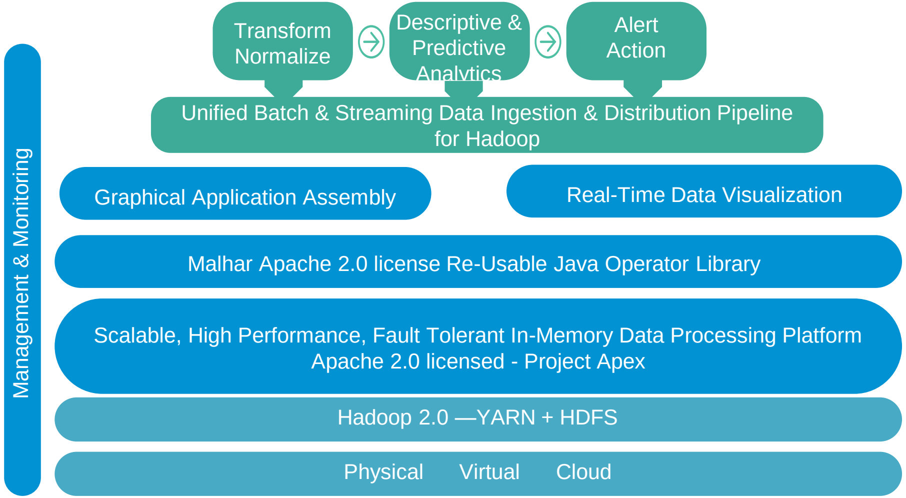
Streaming or Batch Data Processing Visualization

- Intuitive user interface
  - Auto-generate or custom create
  - One dashboard for multiple apps
  - Supports bar, line, pie, area charts & data tables
- Easy to Operate
  - No external component dependency - Runs natively in Hadoop
  - Integrated with DataTorrent management platform
- Simple to extend
  - Any DAG operator can be made a real-time datasource



```
{
  "id": "0.3077799107413739",
  "type": "dataQuery",
  "data": [
    {
      "time": "1429850160000",
      "location": "ALL",
      "advertiser": "ALL",
      "publisher": "ALL",
      "impressions:SUM": "2497900",
      "clicks:SUM": "508742",
      "cost:SUM": "1557925.61",
      "revenue:SUM": "383288.38"
    },
```

Summary



Transform Normalize

Descriptive & Predictive Analytics

Alert Action

Unified Batch & Streaming Data Ingestion & Distribution Pipeline for Hadoop

Graphical Application Assembly

Real-Time Data Visualization

Malhar Apache 2.0 license Re-Usable Java Operator Library

Scalable, High Performance, Fault Tolerant In-Memory Data Processing Platform Apache 2.0 licensed - Project Apex

Hadoop 2.0 —YARN + HDFS

Physical Virtual Cloud

Management & Monitoring

Project Apex

https://www.datatorrent.com/product/project-apex/

DataTorrent RTS Sandbox

https://www.datatorrent.com/download/

- World is moving from ‘Batch’ to ‘Streaming’ BUT both are required
- Need a Hadoop native in memory compute engine that is scalable & fault tolerant in BOTH batch & streaming modes
- With out -of-the box data prep & analytics operators
- Using a consistent & functional development API
- Operationalized through a common management & monitoring layer

DataTorrent

Big Data. Big Actions. Now.

Some Verticals & Use Cases

Ad-Tech

- Real-time customer facing dashboards on key performance indicators
- Click fraud detection
- Billing optimization

Telco & Cable

- Call detail record (CDR) & extended data record (XDR) analysis for
  - Service quality improvement
  - Capacity planning/optimization
- Understanding customer behavior AND context
- Packaging and selling anonymous customer data

Financial Services

- Fraud & risk monitoring
- Sentiment based trading strategies
- Usage based insurance
- Improved credit risk assessment
- Improving turn around time of trade settlement processes

IoT

- Process optimization
- Proactive maintenance prediction
- Remote monitoring & diagnostics