Josh Poduska, Sr. Business Analytics Consultant,

Actian Corporation

Beyond Sampling: Fast, Whole-Dataset Analytics for Big Data on Hadoop

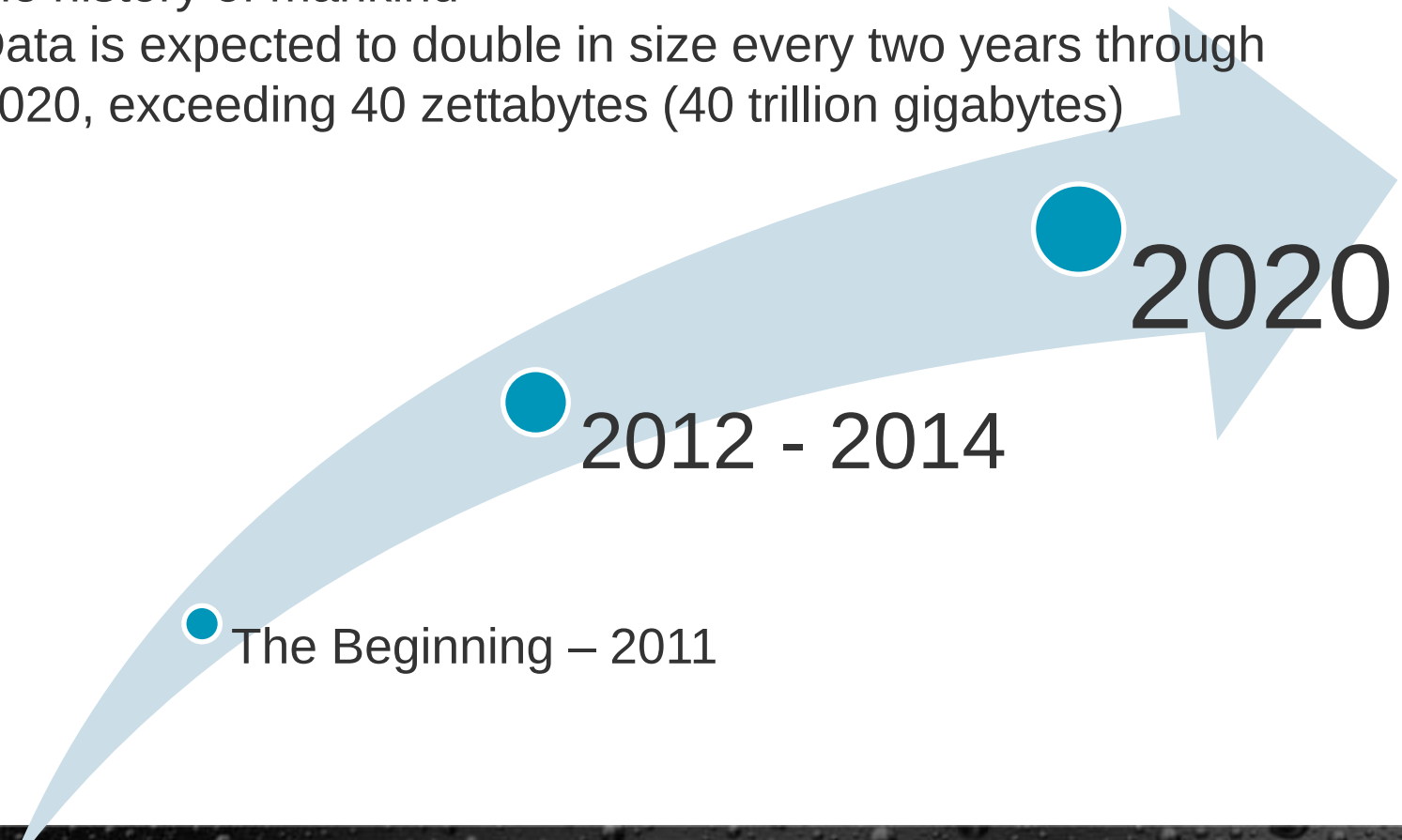
October 2013 KNIME Day Boston

Agenda

- The Age of Data
- Scaling Gap
- Enter the Data Scientists
- POS Hadoop Analytics in KNIME
- Conclusions

The Age of Data

- In the last two years we have generated more data than in the history of mankind
- Data is expected to double in size every two years through 2020, exceeding 40 zettabytes (40 trillion gigabytes)



Entering the Age of Data

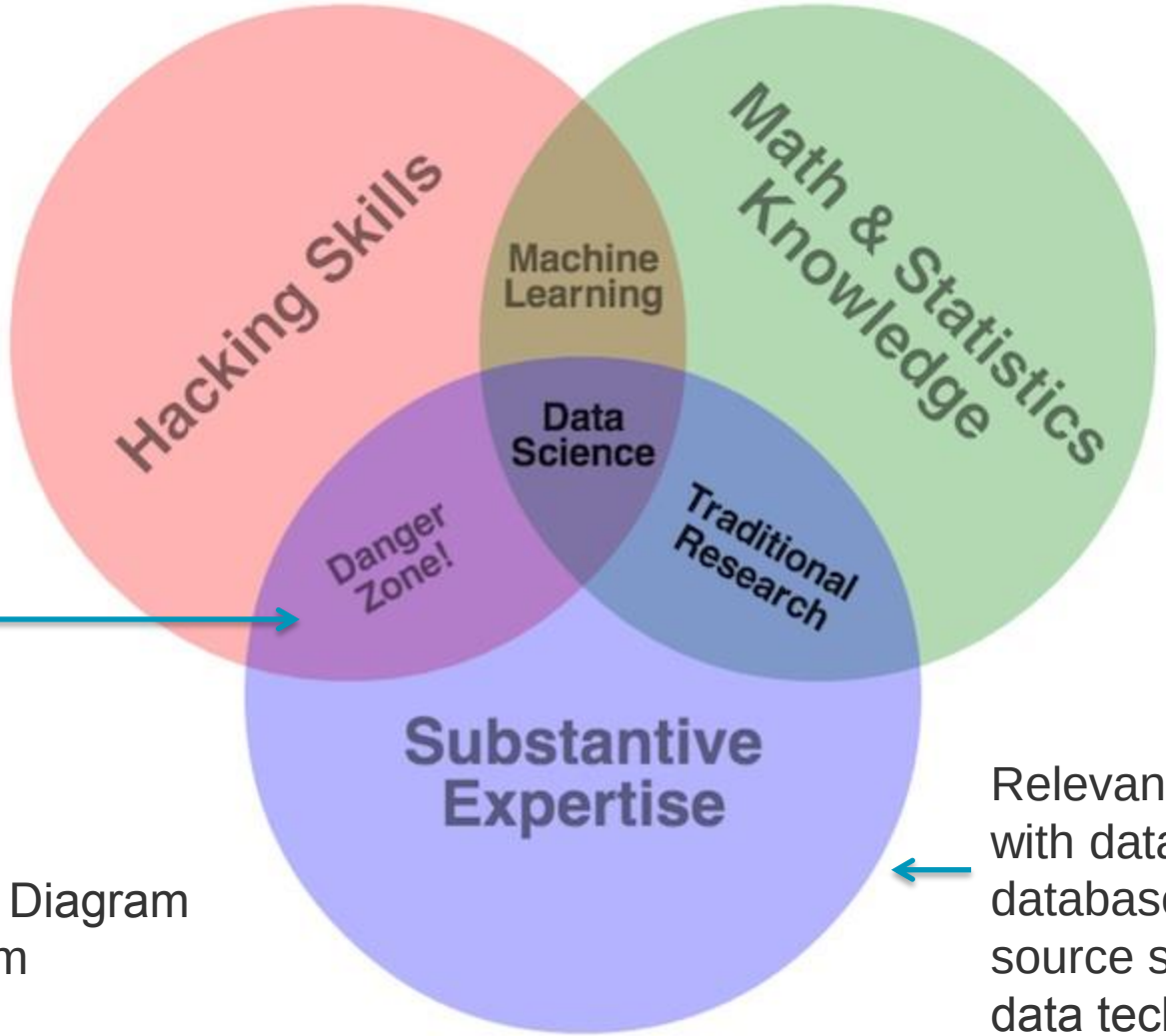
- What's Changed?
  - **Data is THE central business asset:**
    "Data are an organization's sole, non-depletable, non-degrading, durable asset. Engineered right, data's value increases over time because the added dimensions of time, geography, and precision." (Peter Aitken)
  - **Data generation has changed forever**
    - Instrumentation of ALL businesses, people, machines
  - **Data is born digitally and flows constantly**
    - "All things are flowing.." (Heraclitus, 500 BC)

The Scaling Gap

- In the Age of Data, if you are not super-scaling you are failing. What does super-scaling entail?
- Software stacks to **consume**, **analyze** and **act** on event pipelines must be **frictionless** to set up
- Yet extremely **performant:** must **scale-up and scale-out (SUSO)** to fully exploit game-changing price/performance on modern commodity hardware
- And be **elastic**
- And still be **affordable**
- The **hard truth:**
  - Almost no legacy data/event processing stacks super-scale
  - And there is no path to reasonably (and economically) get there

Your legacy analytic software WILL fail in the Age of Data

Venn Diagram of Data Scientists



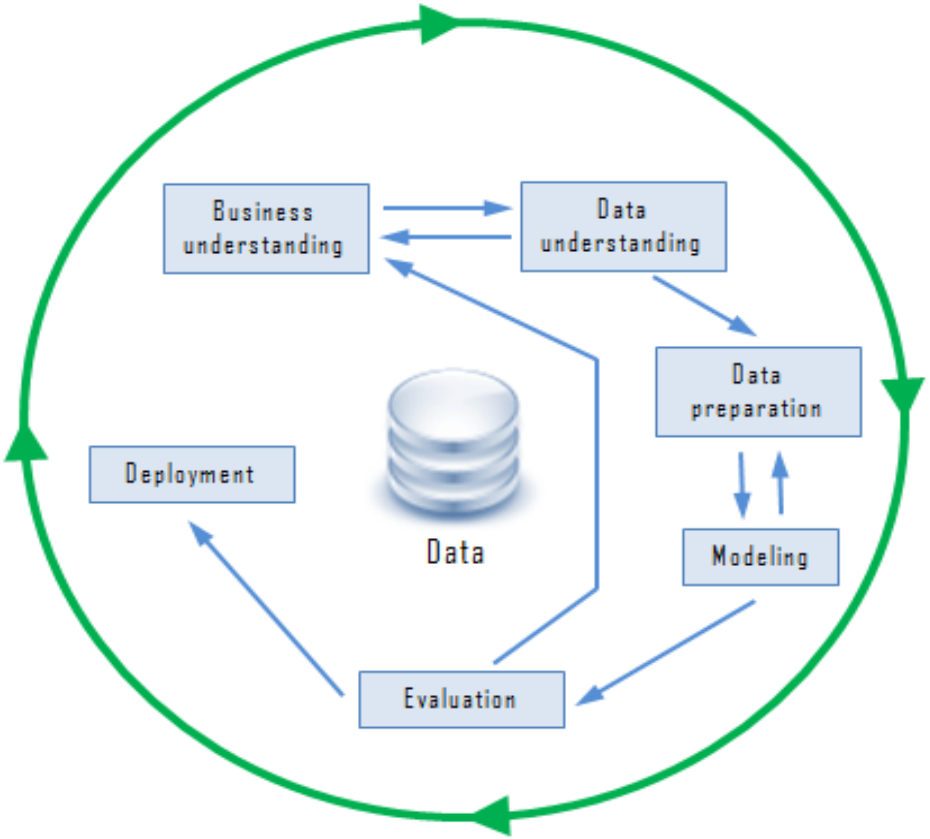
Not statistically valid, lacking interpretation

Drew Conway's Diagram
drewconway.com

Relevant experience with data, tools, databases, open source software, data techniques, …

Data Mining is a Process

- Successfully completing the process requires having a CONVERSATION with the data.



Tools Available (in big data context)

- Distributed Platforms
  - Hadoop, Mesos, …
  - Dataflow, HPCC
- Analytic Databases
  - ParAccel, Vectorwise, Vertica, Aster Data, GreenPlum, Netezza, …
  - Cassandra , HBase, MongoDB, …
- Analytic Platforms
  - Dataflow Analytics, SAS, SPSS, …
  - KNIME, R, RapidMiner, …

Hadoop

- **Open source, distributed (scale out) platform for data processing on cheap hardware**
- Components
  - HDFS – Hadoop Distributed File System
  - MapReduce – computation framework
    - Broken into two phases
    - Map – takes input, produces a name/value mapping
    - Reduce – applies a final reduction of the mapping phase
  - HBase – name/value pair data store
  - Oozie, ZooKeeper, …
  - Distributions: Apache, Cloudera, HortonWorks, MapR, Intel, …

Retail POS Application

Market Basket Analysis

- The Data
  - Retailers have Point of Sales (POS) data
  - Items purchased in same basket are captured (line items)
  - Summary of each basket (basket or order)
  - Information about items - UPC, SKU, description, category hierarchy
- Analysis
  - Need to sell longer-held produce, labeled CLOSEOUT ITEMS
  - Want to know what drives total receipt spend
  - Want to know what items sell well together
  - Which items drive purchase of other items

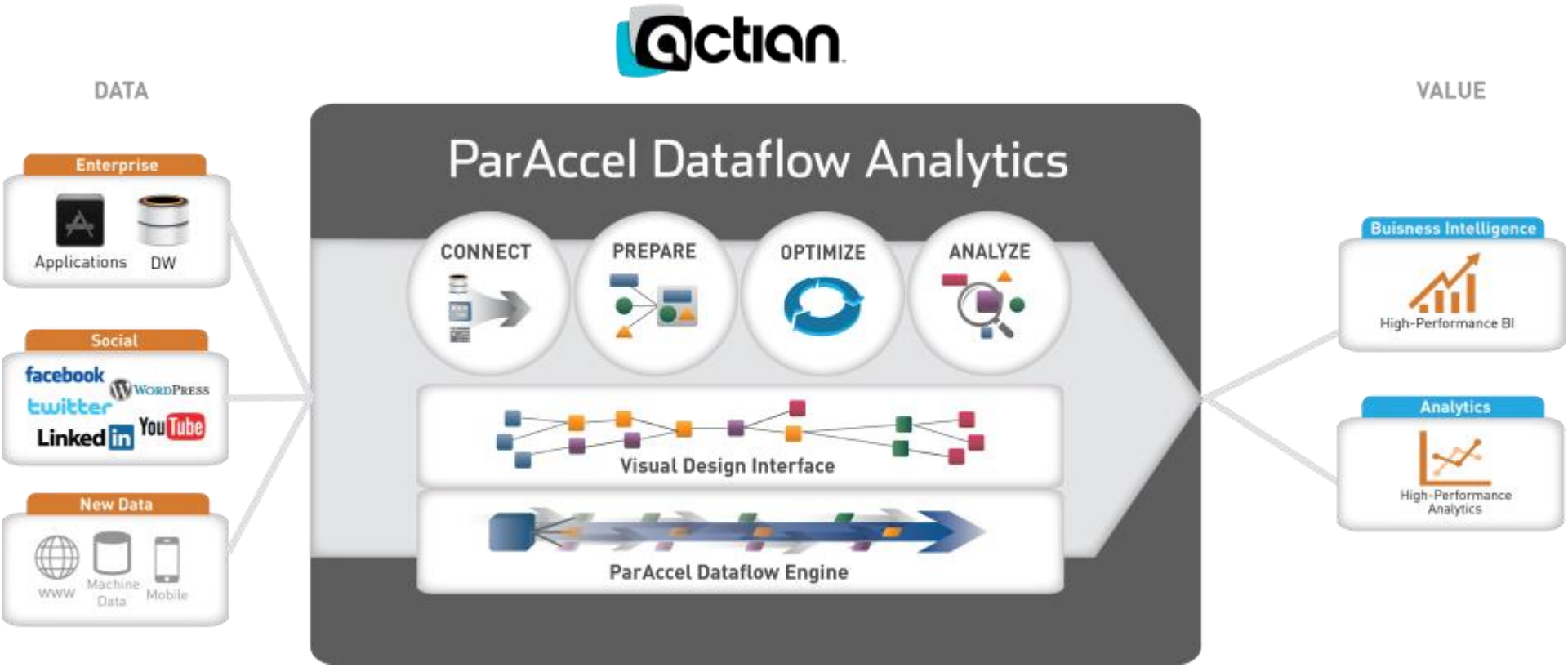
Configuration

- Hundreds of millions of rows of POS data
- Hadoop platform
  - 3 worker nodes, 1 head node
  - Running Cloudera CDH4
  - Distributed Dataflow 6.1
- Analysis Tools (run on desktop)
  - Actian Dataflow Analytics for Hadoop (KNIME installation library)
  - Gephi (open source graph visualization)

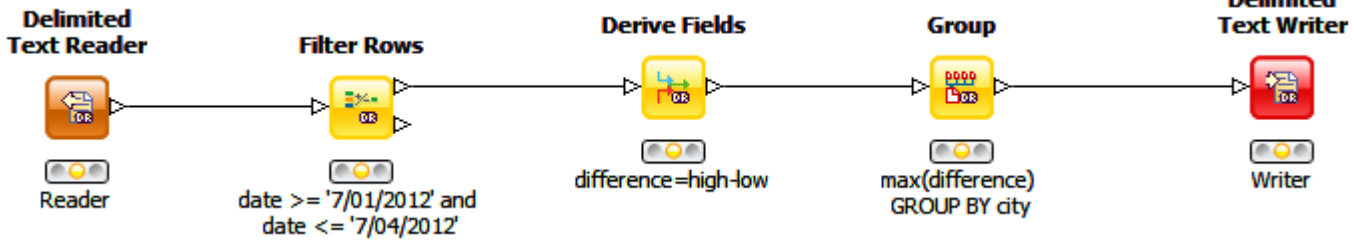
Dataflow Analytics for KNIME

- KNIME (knime.org)
  - Open source analytics workflow platform
  - Highly extensible
  - Active community of plugin contributors
  - Commercially available Server, Teamspace and Report products
- Dataflow Analytics
  - Actian developed extensions to KNIME
  - Includes scalable Dataflow technology
  - Large set of “nodes” based on Dataflow

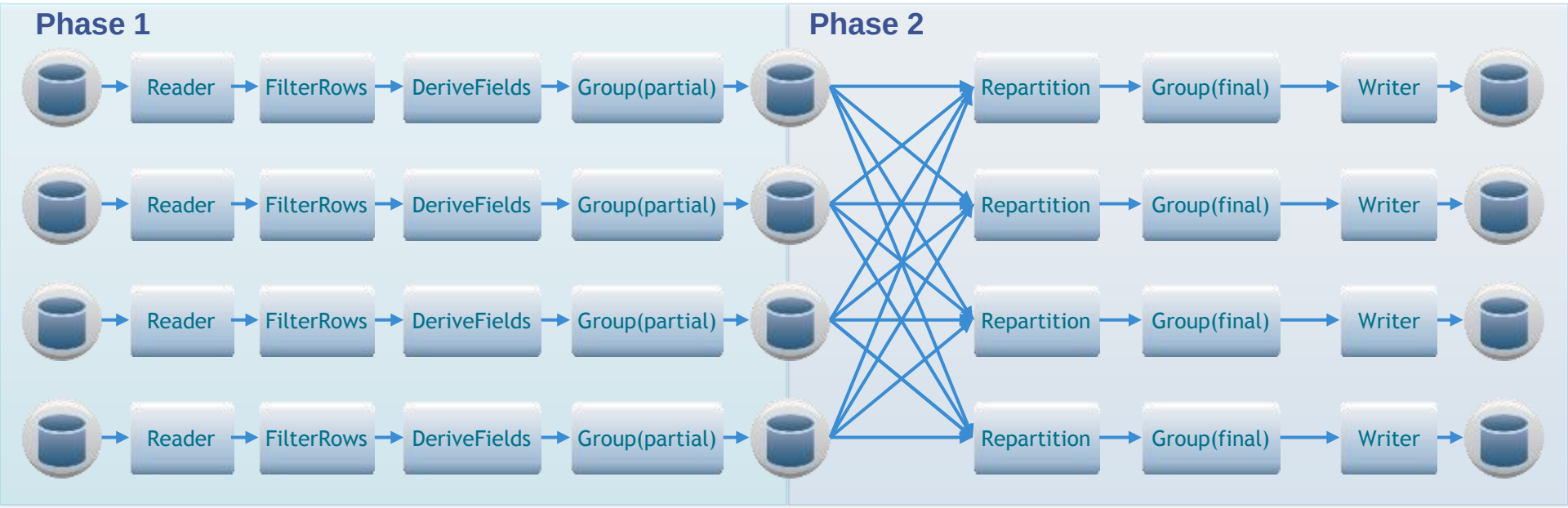
Dataflow Analytics for KNIME



Dataflow Analytics for KNIME

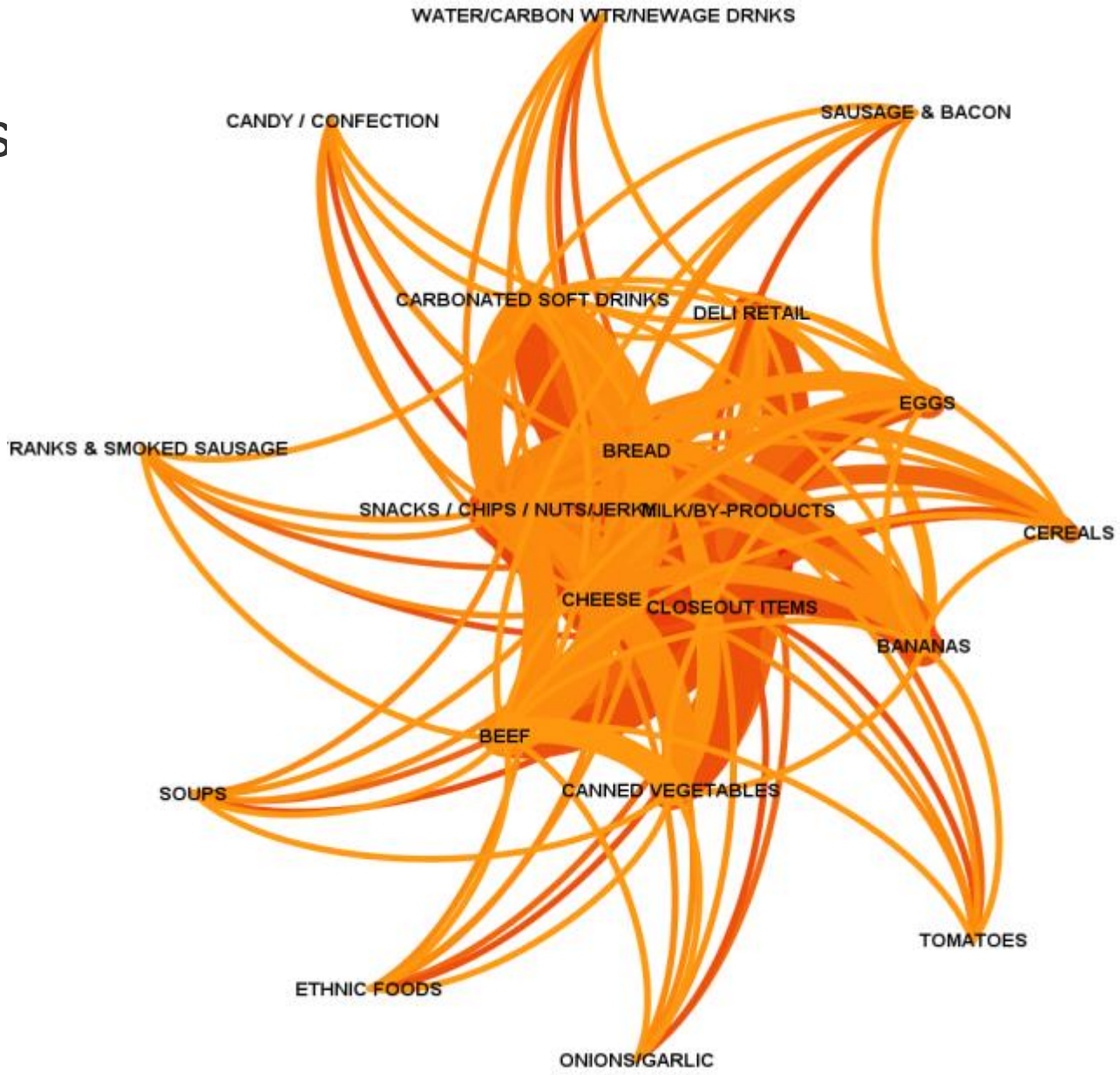


Compiled to a set of physical graphs



Demo

Visualize Associations in Gephi



Big Data Conversation → Analytical Impact

- Retail Goal: Increase Spend on Closeout Produce

- High confidence antecedent:
  - Cheese + Bananas
  - Snacks + Bananas
  - Bread + Bananas …
- High support antecedents:
  - Bananas
  - Tomatoes
  - Milk + Cheese …

Big Data Conversation → Analytical Impact

- Retail Goal: Increase Spend on Closeout Produce

- What Closeout Produce associations exist with low lift? (↑ lift → bring consequent into basket)
  - Chicken or Cheese + Soft Drinks
  - ➢ **Action = Same visit coupon**

Big Data Conversation → Analytical Impact

- Retail Goal: Increase Spend on Closeout Produce

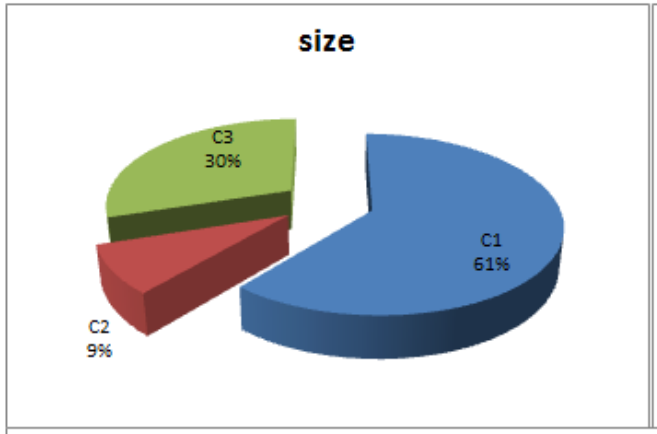
- What has high support but no Closeout Produce association?
  - Candy, Canned Vegetables, Soup.
  - ➢ **Action = Same visit coupon, receipt coupon, delivered coupon, etc.**

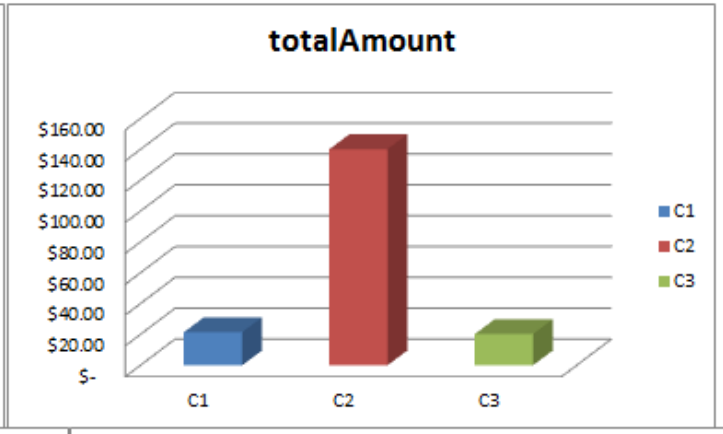
Big Data Conversation → Analytical Impact

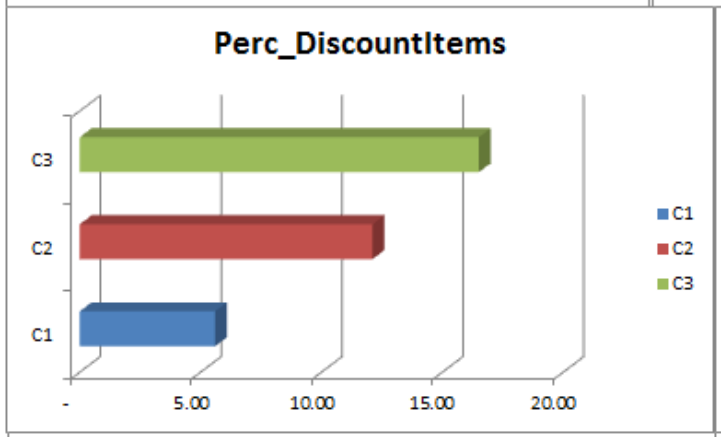
- Retail Goal: Increase Total Spend per Customer

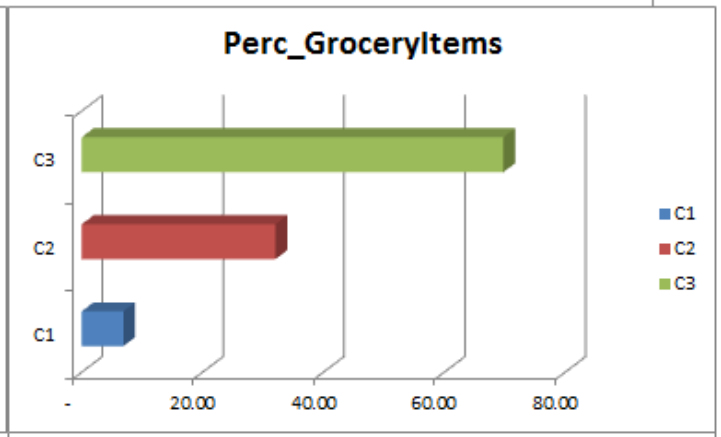
- Clustering points to % Basket in Market Department being a key factor in total spend.

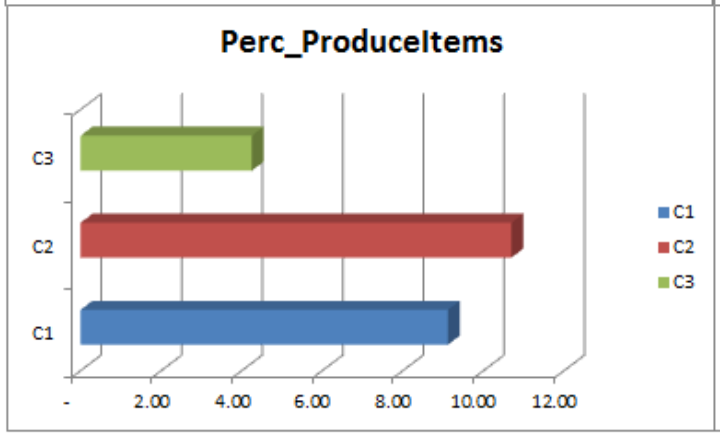
Cluster Results

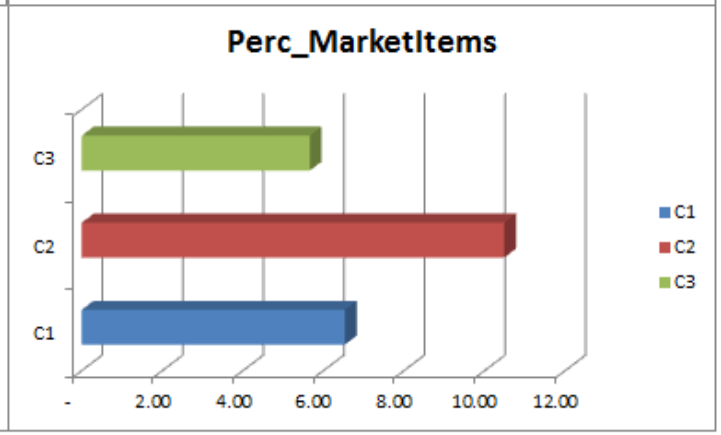












Big Data Conversation → Analytical Impact

- Retail Goal: Increase Total Spend per Customer

- Follow up Cluster with Linear Regression
  - % Basket in Market Department is indeed a key predictor even after factoring for:
    - **DOW**
    - **Season**
    - **%Basket Market Department**
    - **%Basket Produce Department**
    - **%Basket Grocery Department**
    - **%Basket Discount Department**

Conclusion

- The Age of Data is here
  - Data is the central business asset
  - Data generation has changed forever
  - Shift of analysis focus to time-stamped events
  - Crisis of software that scales to meet demand
- Data Science is changing how data is:
  - Collected, discovered, analyzed, used, acted upon …
- Big Data Conversations
  - Deep analysis is required to move beyond basic findings
  - Actionable results require very heavy lifting

Questions

www.actian.com

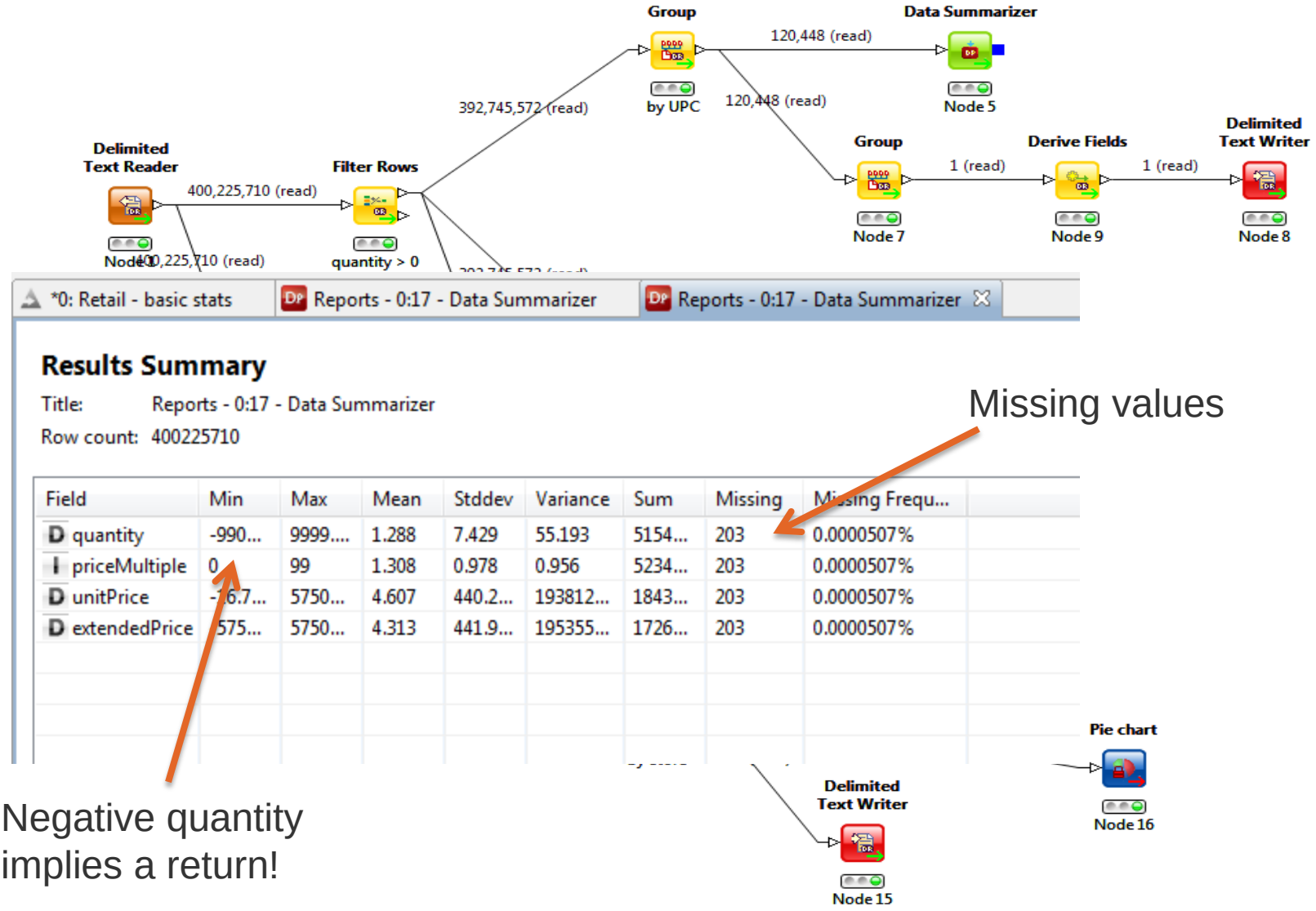
facebook.com/actiancorp

@actiancorp

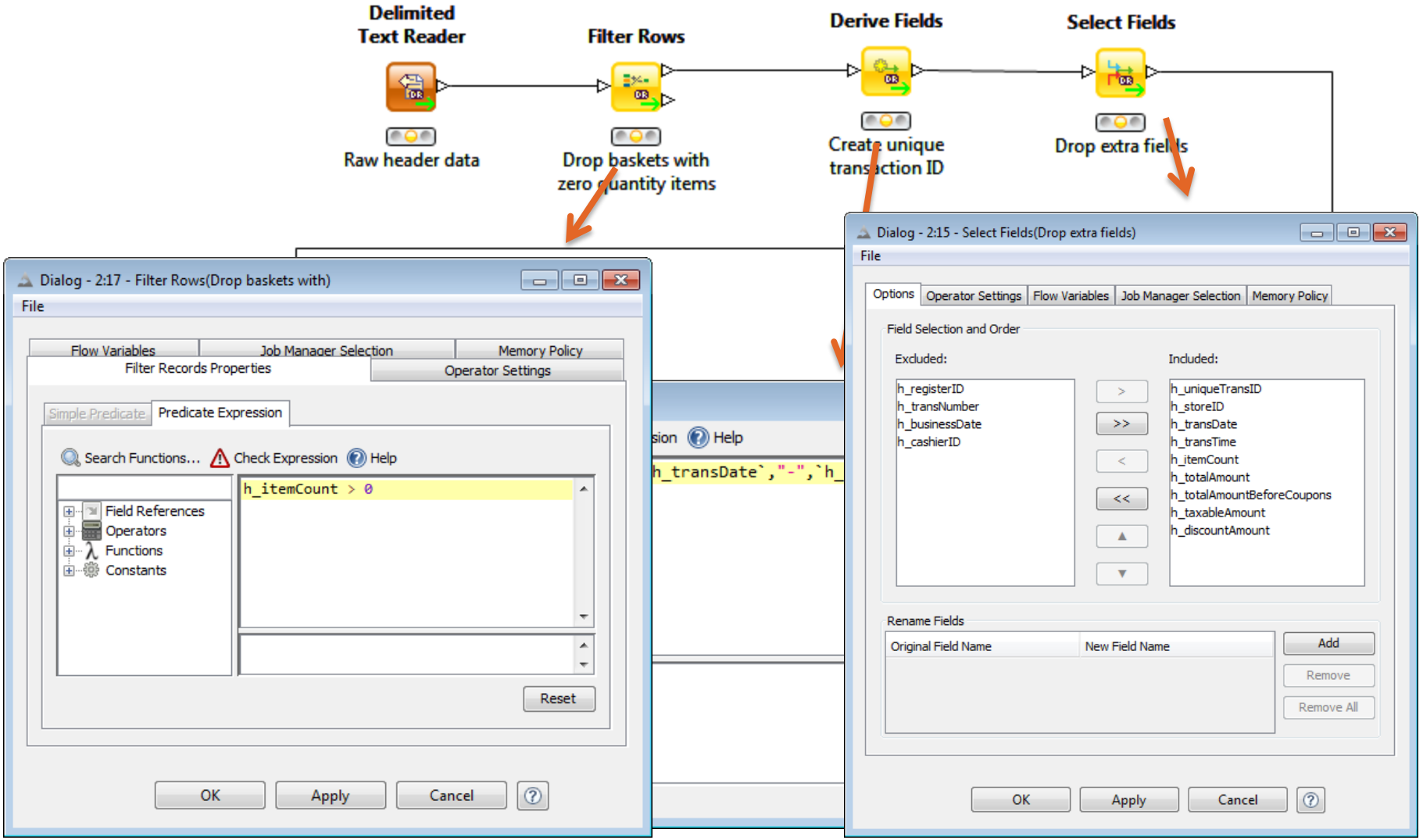
joshua.poduska@action.com

Backup slides

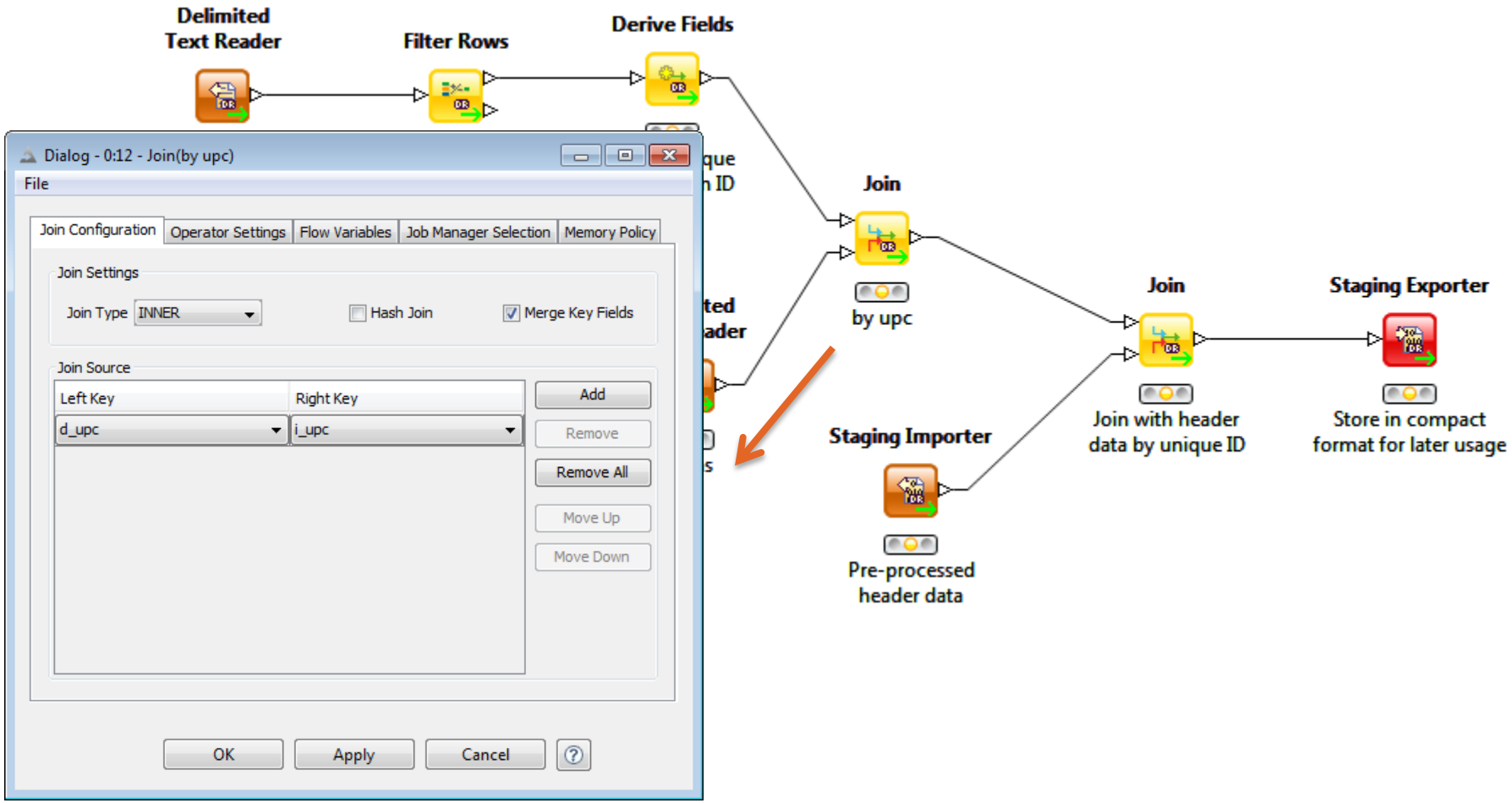
Start with data discovery



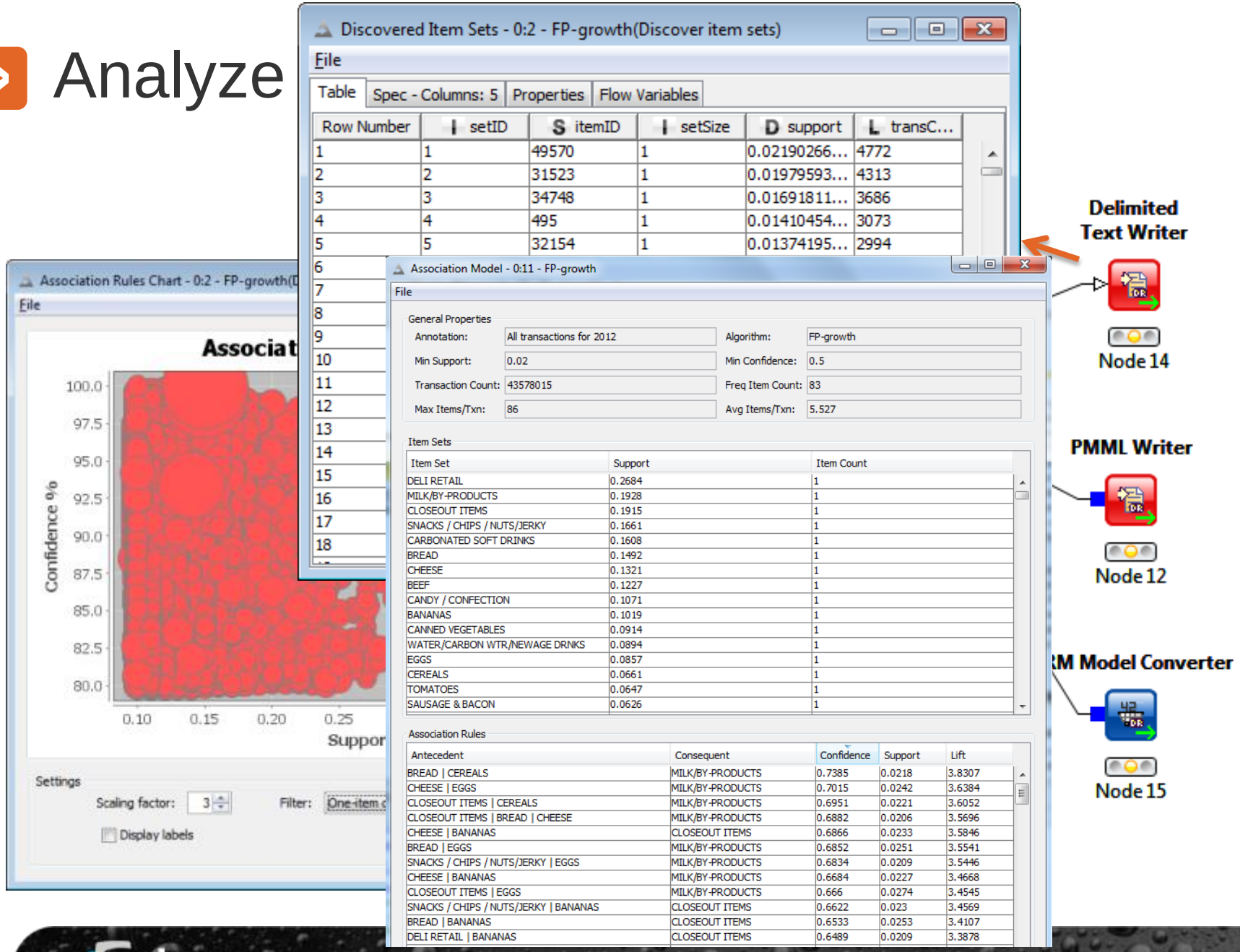
Cleanse, enrich & aggregate



Cleanse and enrich

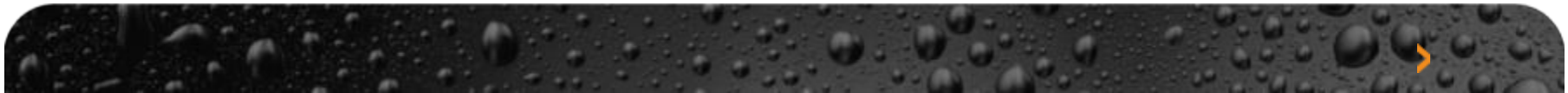


Analyze


Discovered Item Sets - 0:2 - FP-growth(Discover item sets)
File
Table
Spec - Columns: 5
Properties
Flow Variables
Row Number | setID | itemID | setSize | support | transC...
1 | 1 | 49570 | 1 | 0.02190266... | 4772
2 | 2 | 31523 | 1 | 0.01979593... | 4313
3 | 3 | 34748 | 1 | 0.01691811... | 3686
4 | 4 | 495 | 1 | 0.01410454... | 3073
5 | 5 | 32154 | 1 | 0.01374195... | 2994
Association Rules Chart - 0:2 - FP-growth(D
File
Confidence %
100.0
97.5
95.0
92.5
90.0
87.5
85.0
82.5
80.0
0.10
0.15
0.20
0.25
Settings
Scaling factor:
3
Filter:
Display labels
Association Model - 0:11 - FP-growth
File
General Properties
Annotation: All transactions for 2012
Algorithm: FP-growth
Min Support: 0.02
Min Confidence: 0.5
Transaction Count: 43578015
Freq Item Count: 83
Max Items/Txn: 86
Avg Items/Txn: 5.527
Item Sets
Item Set | Support | Item Count
DELI RETAIL | 0.2684 | 1
MILK/BY-PRODUCTS | 0.1928 | 1
CLOSEOUT ITEMS | 0.1915 | 1
SNACKS / CHIPS / NUTS/JERKY | 0.1661 | 1
CARBONATED SOFT DRINKS | 0.1608 | 1
BREAD | 0.1492 | 1
CHEESE | 0.1321 | 1
BEEF | 0.1227 | 1
CANDY / CONFECTION | 0.1071 | 1
BANANAS | 0.1019 | 1
CANNED VEGETABLES | 0.0914 | 1
WATER/CARBON WTR/NEWAGE DRNKS | 0.0894 | 1
EGGS | 0.0857 | 1
CEREALS | 0.0661 | 1
TOMATOES | 0.0647 | 1
SAUSAGE & BACON | 0.0626 | 1
Association Rules
Antecedent | Consequent | Confidence | Support | Lift
BREAD | CEREALS | MILK/BY-PRODUCTS | 0.7385 | 0.0218 | 3.8307
CHEESE | EGGS | MILK/BY-PRODUCTS | 0.7015 | 0.0242 | 3.6384
CLOSEOUT ITEMS | CEREALS | MILK/BY-PRODUCTS | 0.6951 | 0.0221 | 3.6052
CLOSEOUT ITEMS | BREAD | CHEESE | MILK/BY-PRODUCTS | 0.6882 | 0.0206 | 3.5696
CHEESE | BANANAS | CLOSEOUT ITEMS | 0.6866 | 0.0233 | 3.5846
BREAD | EGGS | MILK/BY-PRODUCTS | 0.6852 | 0.0251 | 3.5541
SNACKS / CHIPS / NUTS/JERKY | EGGS | MILK/BY-PRODUCTS | 0.6834 | 0.0209 | 3.5446
CHEESE | BANANAS | MILK/BY-PRODUCTS | 0.6684 | 0.0227 | 3.4668
CLOSEOUT ITEMS | EGGS | MILK/BY-PRODUCTS | 0.666 | 0.0274 | 3.4545
SNACKS / CHIPS / NUTS/JERKY | BANANAS | CLOSEOUT ITEMS | 0.6622 | 0.023 | 3.4569
BREAD | BANANAS | CLOSEOUT ITEMS | 0.6533 | 0.0253 | 3.4107
DELI RETAIL | BANANAS | CLOSEOUT ITEMS | 0.6489 | 0.0209 | 3.3878
Delimited Text Writer
Node 14
PMML Writer
Node 12
M Model Converter
Node 15

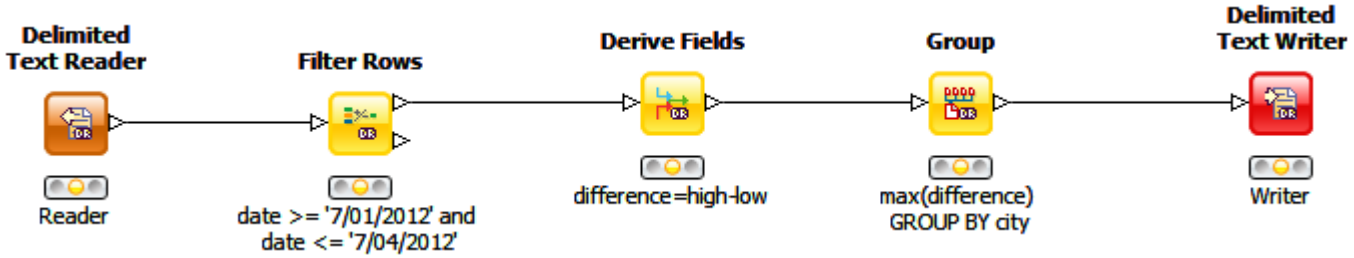
History of Dataflow

- Initially developed as next-gen data engine for integration
- Used to be DataRush
- Requirements
  - High data throughput
  - Scalable (data, multicore)
  - Based on dataflow concepts
  - Component based architecture
  - Easy to extend
  - Easily fits in visual development environment
- Embedded in Pervasive products (DataProfiler)
- Extended with SDK for more general use

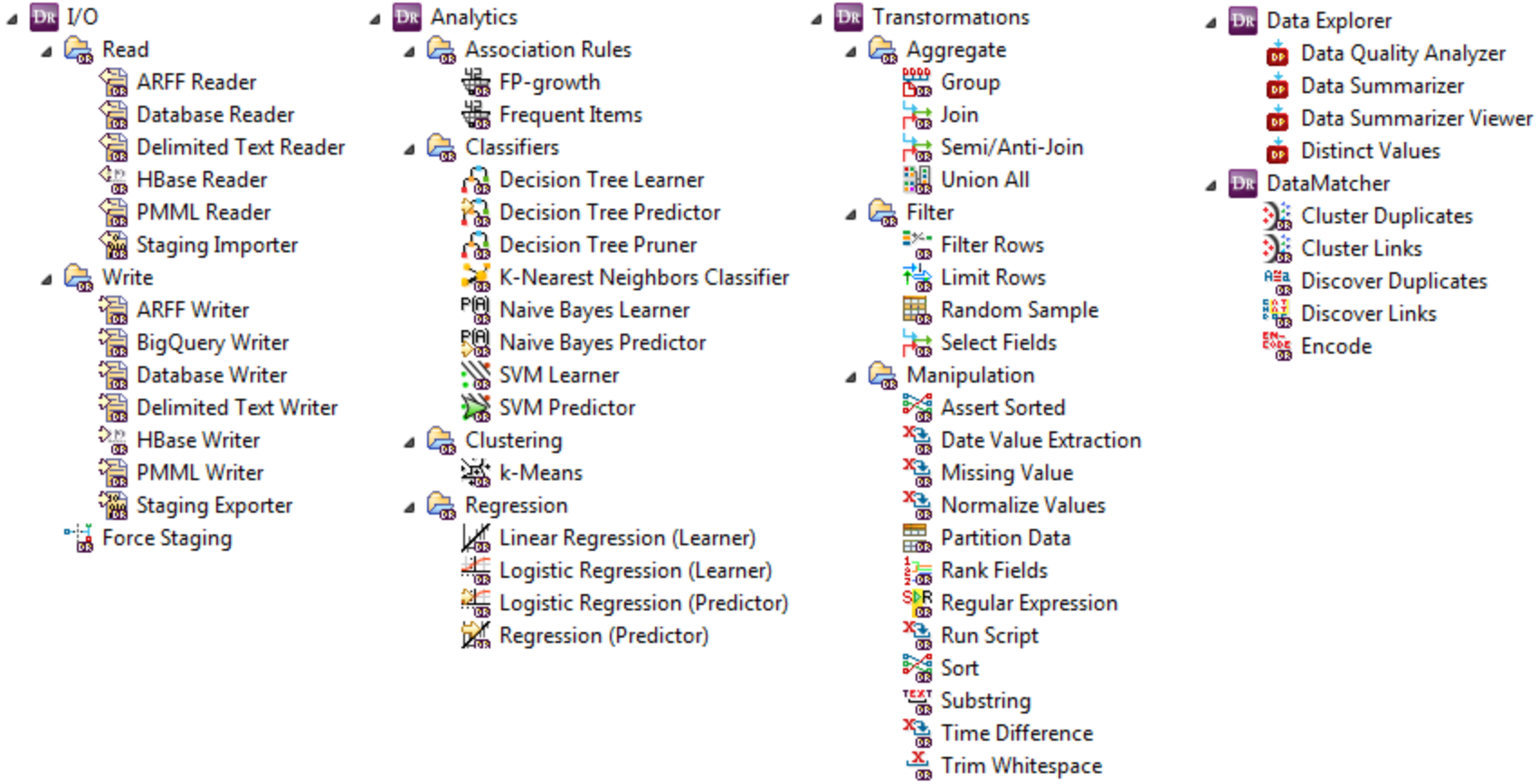


Dataflow Concepts

- Operators (nodes) linked together in a directed graph
- Data flows along edges
- Shared nothing architecture
- Provides pipeline parallelism
- Supports data parallelism
- Data scalable

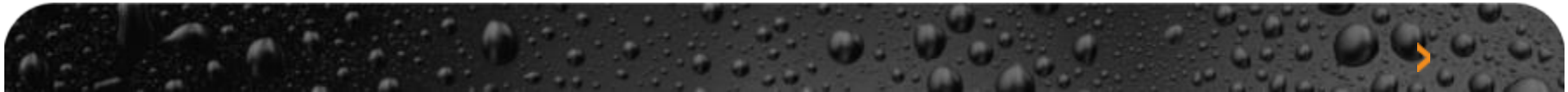


Operator Library

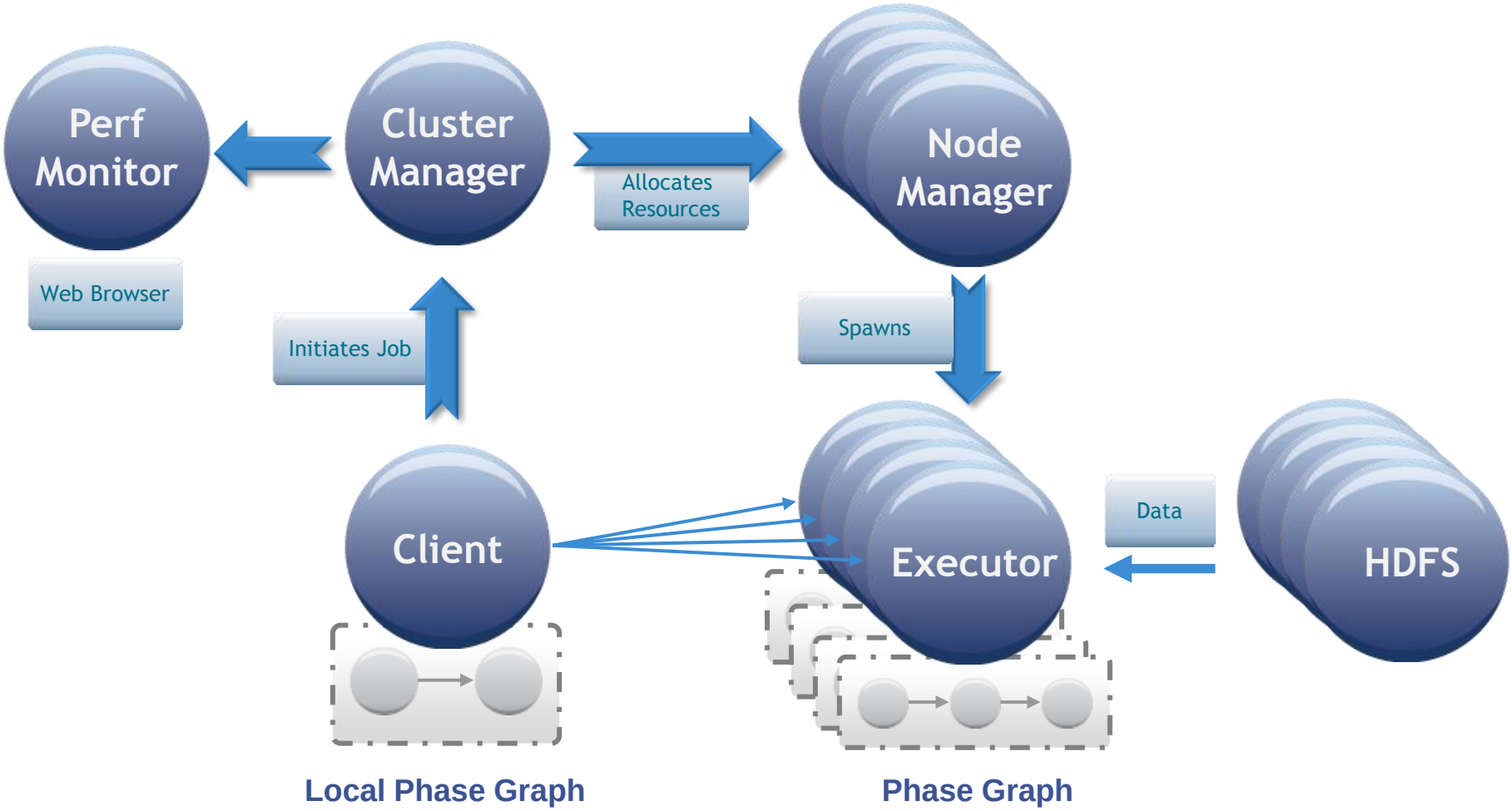


Integration with Hadoop

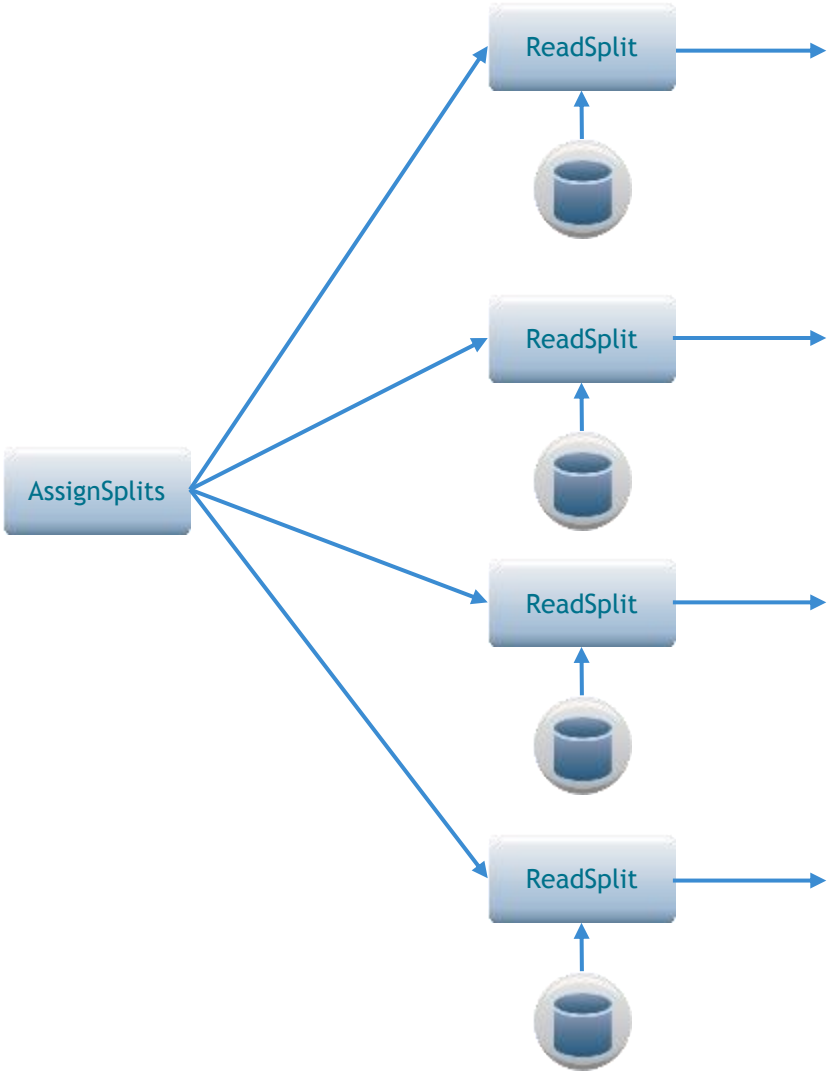
- Data Level
  - HDFS access
    - File system abstraction – works with all I/O operators
    - Distributed execution – uses splits much like MR
  - HBase
    - Temporal key-value data store based on column families
    - Fast loading using HFile integration
    - Fast temporal queries
- Execution
  - Distributed execution uses distribute DataRush engines (not MapReduce)
  - Integrating with YARN for resource sharing



Distributed Execution



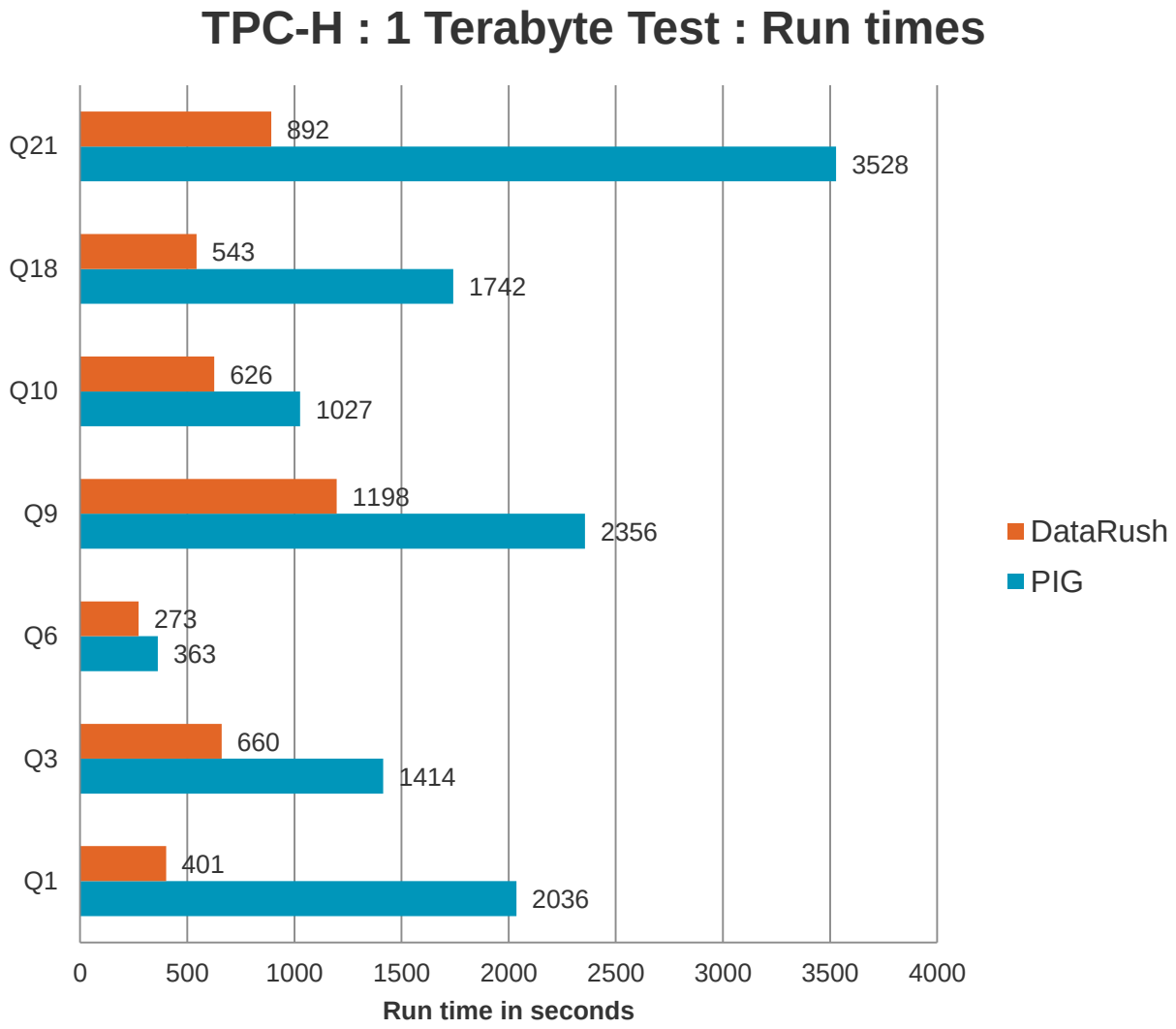
Distributed I/O



- Allows downstream operators to be parallelized
- Parallelization concepts are the same whether the graph is run locally or distributed

Performance Test

- Dataflow versus PIG
  - Used TPC-H data
  - Generated 1TB data set in HDFS
  - Ran several “queries” coded in Dataflow and PIG
  - Run times in seconds (smaller is better)



Cluster Configuration:

- 5 worker nodes
- 2 X Intel E5-2650 (8 core)
- 64GB RAM
- 24 X 1TB SATA 7200 rpm

Dataflow Analytics Solutions

- Opera Solutions
  - Data science solutions provider
  - Embedding DataRush in engineered solutions
- Healthcare
  - Claims cleansing & processing
- Retail
  - Market basket analysis
  - Product category resolution (MDM)
- Telecom
  - CDR processing & analysis

"[Dataflow's] efficiency and ability to automatically scale, whether on a single server or a Hadoop cluster, supports our vision for consistent, reusable, scalable Big Data analytics."

– Armando Escalante, Chief Operating Officer, Opera Solutions