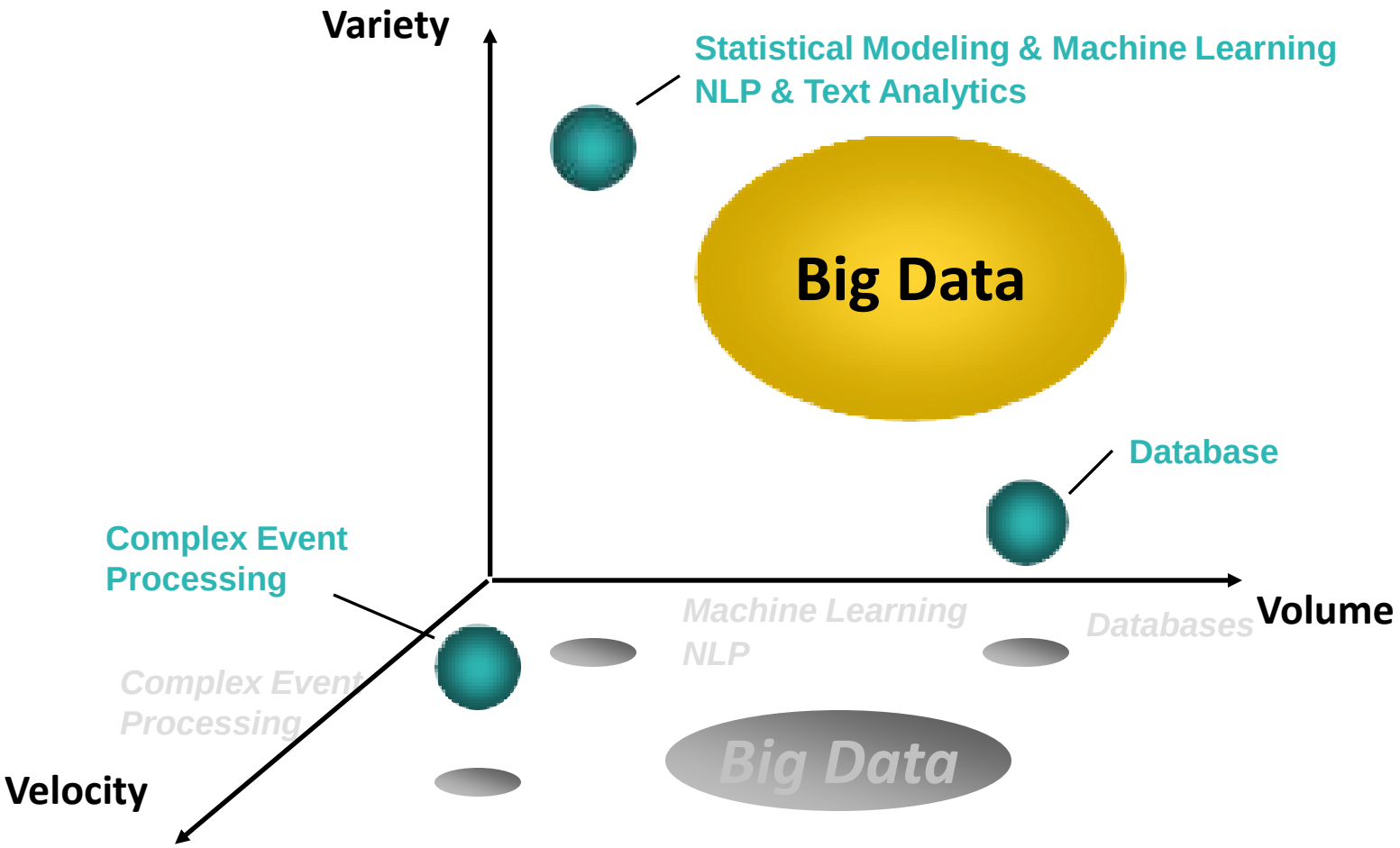
Big Data Analytics – Enterprise Opportunities and Case Studies

Shivakumar Vaithyanathan
IBM Chief Scientist – Text Analytics
Department Head – Analytics, IBM Research, Almaden

Outline

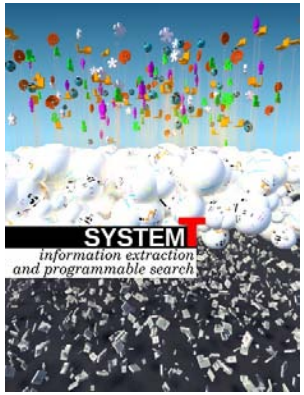
- Changing world of Enterprise Applications
- Case Studies
  - Digital Marketing Effectiveness – Media & Entertainment
  - Customer Retention and New Customer Acquisition – Retail Banking
  - System Log Analysis in a Data-Center – Cross-Industry
  - Slicing and dicing of Sensor and Simulation Data – Renewable Energy
  - Enterprise Knowledge Management and Search – Cross-Industry
- Quick Overview of Big Data Analytics Tools

Big Data vis-à-vis existing infrastructure and analytics



It is Really Big Data and Complex Analytics

Text

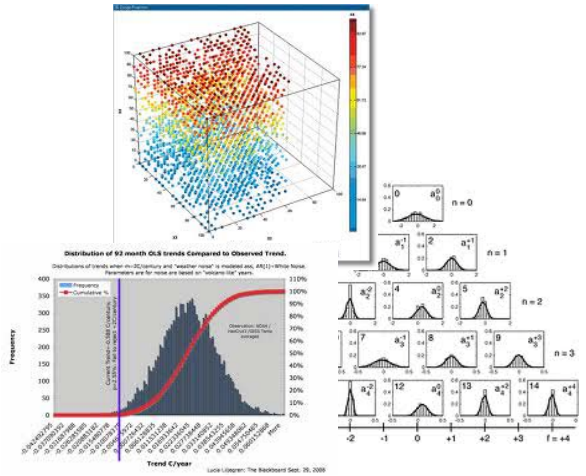


Unstructured Analytics

Entity Resolution
Record Linkage

Clickstream Data

Logs &
Transactions



Statistical Modeling and Predictive Analytics

The Changing World of Enterprise Applications

Enterprise IT Applications

- Log analytics and event monitoring
- Enterprise Knowledge Management
  - Contact centre management
  - Enterprise Knowledge Management & Discovery
  - Compliance and eDiscovery

Line of Business Applications

- Customer retention and new customer acquisition
- Digital marketing effectiveness
  - Reputational risk
  - Campaign management
- Decision support for investment advisors

Outline

- Changing world of Enterprise Applications
- Case Studies
  - Digital Marketing Effectiveness – Media & Entertainment
  - Customer Retention and New Customer Acquisition – Retail Banking
  - System Log Analysis in a Data-center – Cross-Industry
  - Slicing and dicing of Sensor and Simulation Data – Renewable Energy
- Quick Overview of Big Data Analytics Tools

Case Study: Effectiveness of an Ad Campaign for a Movie Studio

Film A

Film B

How many people are talking about the film ?

- Do they intend to actually see the film?
- Did the trailers have any impact?

Who are they ?

- What is their demographic profile
- Are they highly influential?
- Are they avid movie-goers?
- Are they comic book fans?

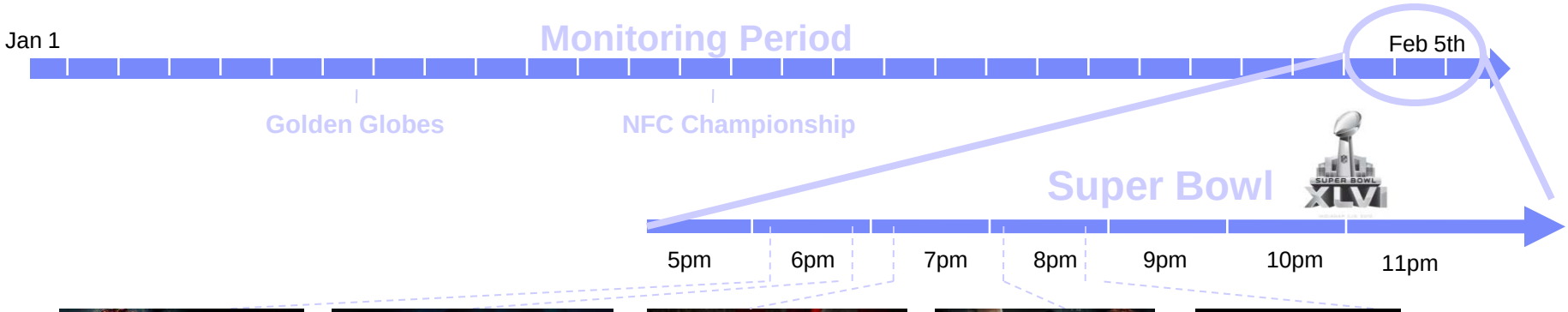
What is their reaction?

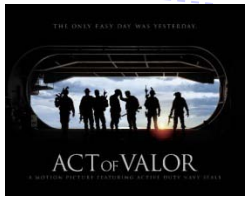
- Did they like the trailer?
- What elements (plot, characters, etc.) had the best reaction?
- What elements (plot, characters, etc.) had the worst reaction?
- Why did they feel this way?

How does this compare to the competition?

- Compared to other trailers aired at the same time?
- Compared to other films releasing at the same time?

To-the-minute insight over a one month period





Data Set

- > 3 B tweets
- 5.7M blog and forum posts
- 3.5M relevant messages

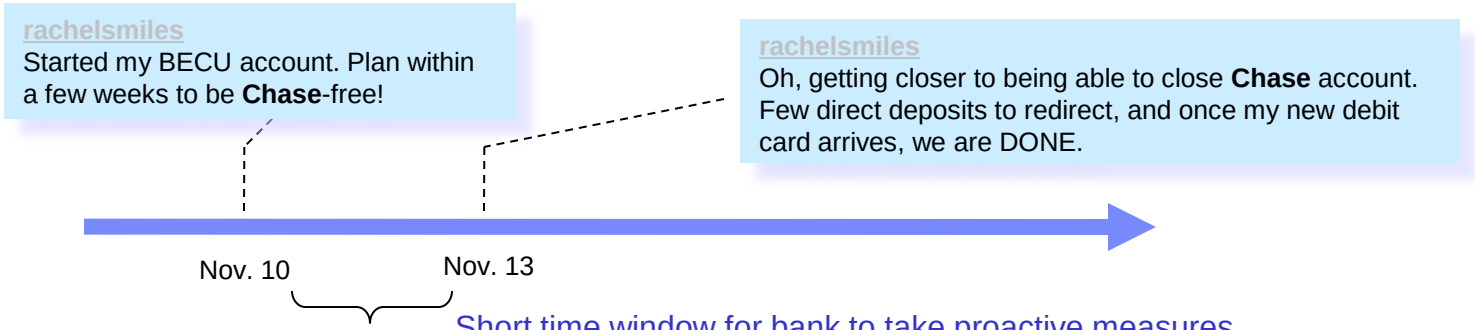
Information extracted

- Buzz, sentiment and intent
- Gender, Location and Occupation
- Avid movie-goers, comic book fans
- Specific attributes of the film/trailer

Case Study : Customer Retention, New Customer Acquisition and Lead Generation

- Customer Retention by identifying individual customers who are likely to churn using explicit and implicit defection triggers
- New Customer Acquisition through targeting of prospects looking for a new service provider or dissatisfied with competitors products and services
- Lead Generation for products and services such as credit card offers, travel discount offers, home mortgage loans and educational loans

Customer Retention Opportunities



New Customer Acquisition : Prospecting Opportunities

... so I am moving my $ from @bankofamerica to a credit union. Recommendations in SF?

Prospective customer for banks in San Francisco

Lead Generation Opportunities

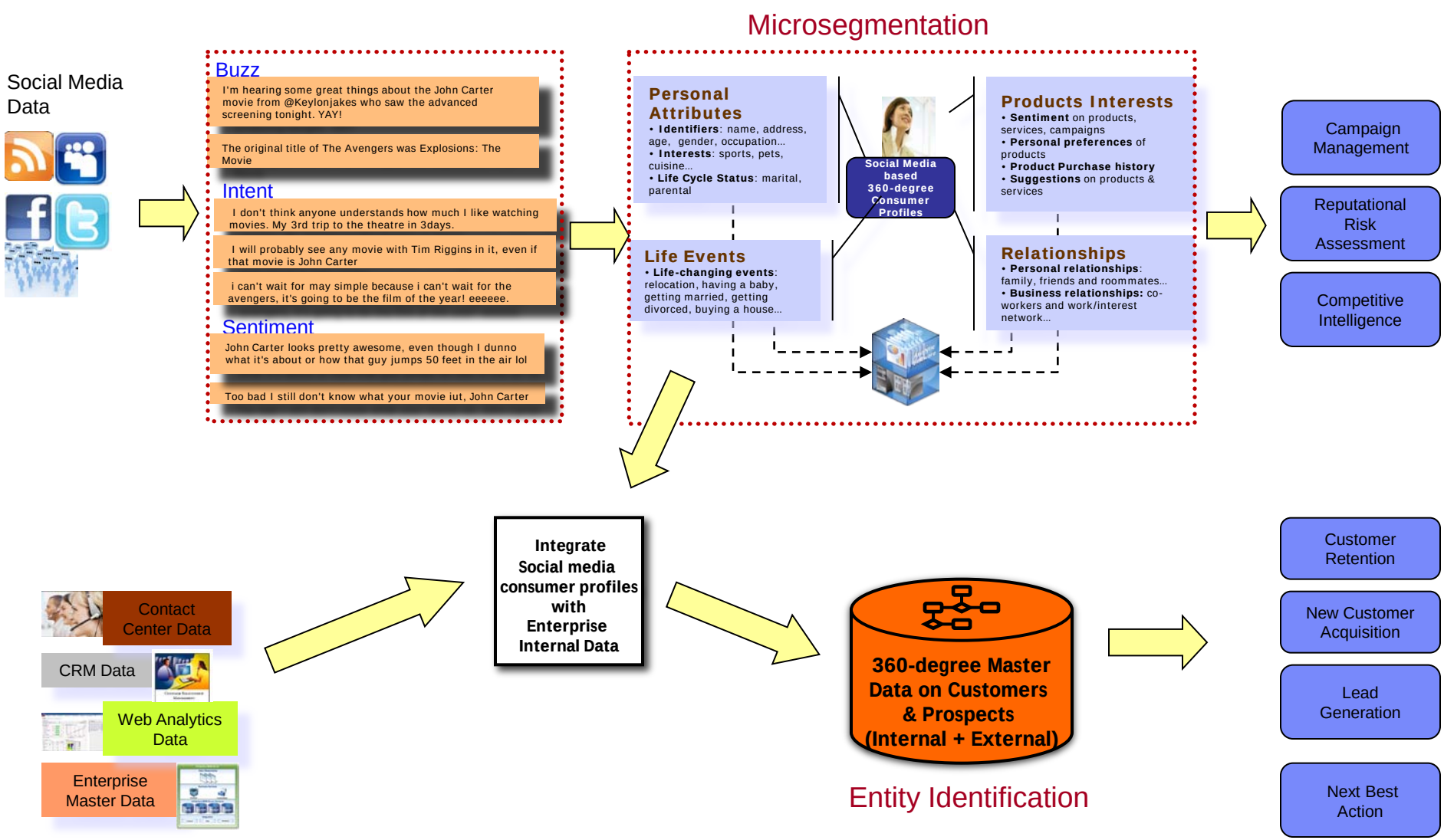
@scarlett_cherry Congratulations! Baby Cherry has finally decided to come into the world! :D Well done!

Credit card offers for baby purchases

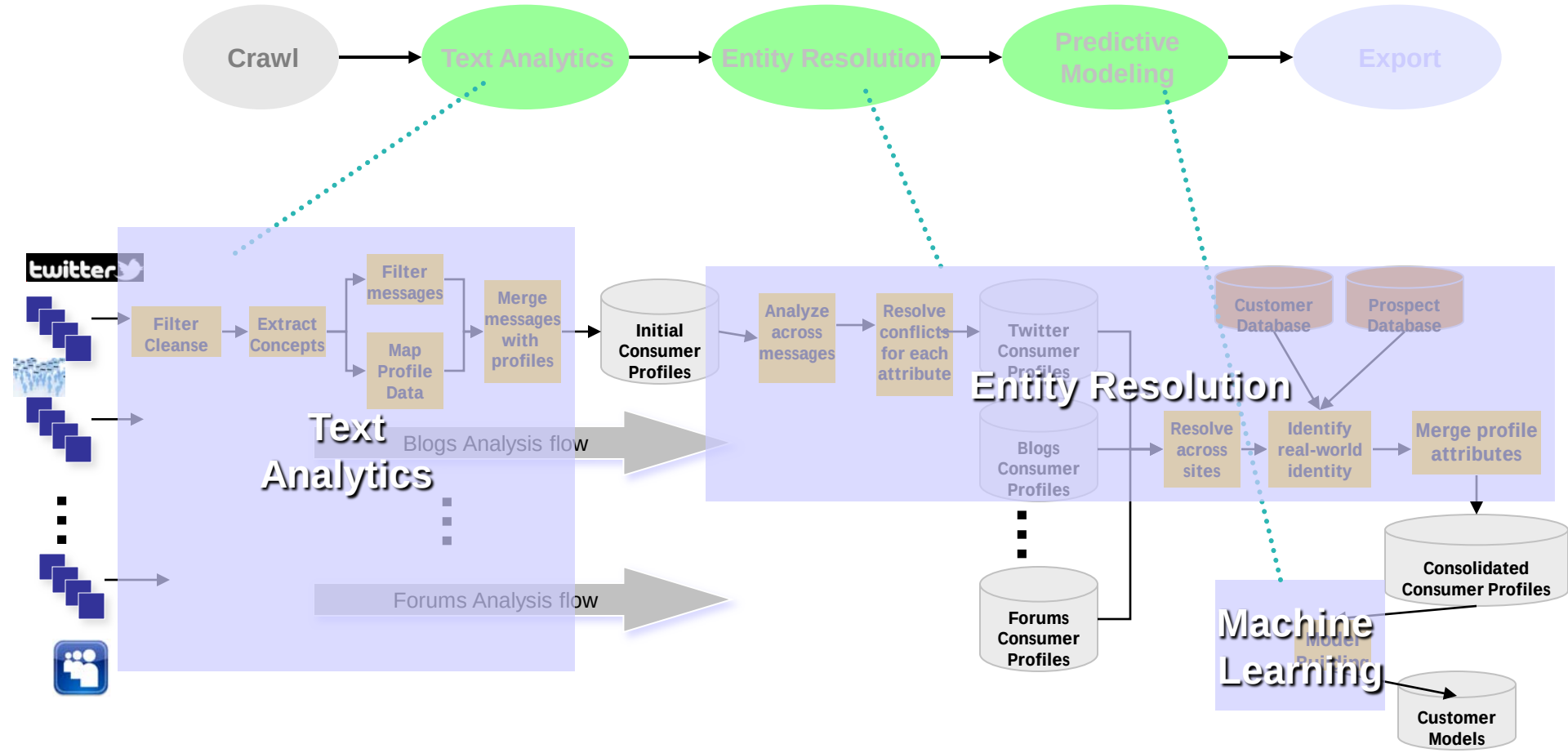
I'm thinking about buying a home in Buckingham Estates per a recommendation. Anyone have advice on that area? #atx #austinrealestate #austin

Home mortgage loans, Credit Card offers

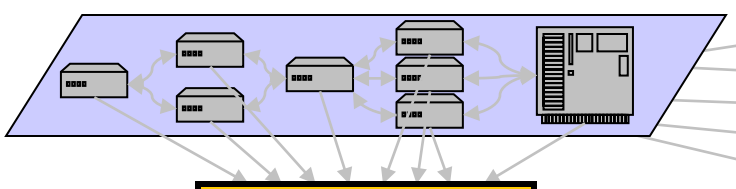
360-degree Profile Management for Retail Banking



360-degree Profile Management



Case Study: Root Cause Analysis of abnormal KPI values in a data center



System Log Collector

System Log data

Performance Data Collector

Performance data

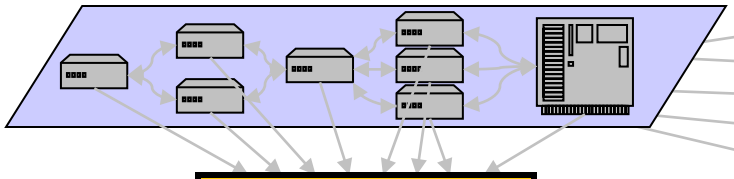
Whenever a performance indicator goes out of the acceptable range...

...identify all events in the system logs that could be related to the abnormal performance.

Machine ID	Time	KPI	Val	
x.x.x.x	1:00	Number of Output Packets Dropped		...
x.x.x.x	1:00	CPU Utilization (%)	20	...
x.x.x.x	2:00	Number of Output Packets Dropped	100000	...
x.x.x.x	2:00	CPU Utilization (%)	22	...
		...	...	...

Fault

Case Study: Root Cause Analysis of abnormal KPI values in a data center



System Log Collector

Parsing & Attribute Value Extraction

Machine	Tim	Message Type	...

Syslog data

Performance Data Collector

Performance data

Record Linkage

Lookup Table

- What KPI values represent faults?
- What syslog message types could be relevant to each type of fault?

Machine ID	Time	KPI	Value	....
x.x.x.x	1:00	Number of Output Packets Dropped	0	...
x.x.x.x	1:00	CPU Utilization (%)	20	...
x.x.x.x	2:00	Number of Output Packets Dropped	100000	...
x.x.x.x	2:00	CPU Utilization (%)	22	...
		...	...	...

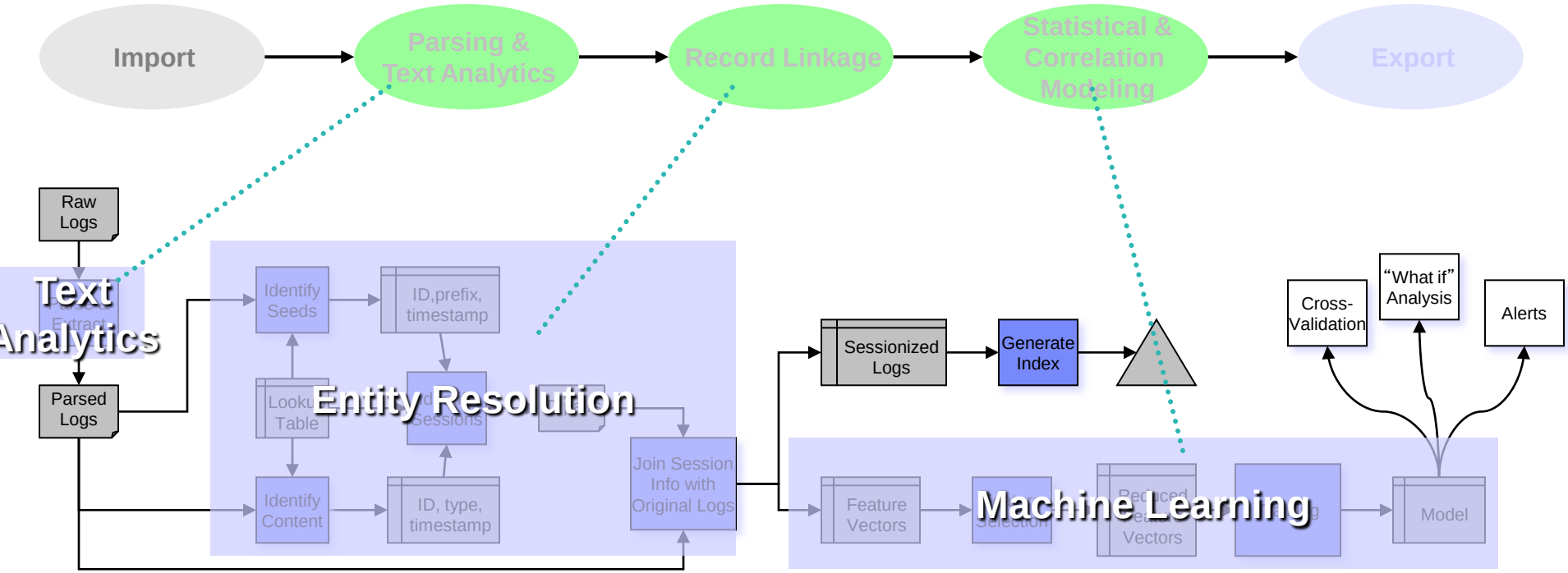
Fault

Data Transformation
Session identification,
Negative example generation,
Feature vector generation

Advanced Analytics
Correlation and
Root Cause Analysis

Log Analytics Flow

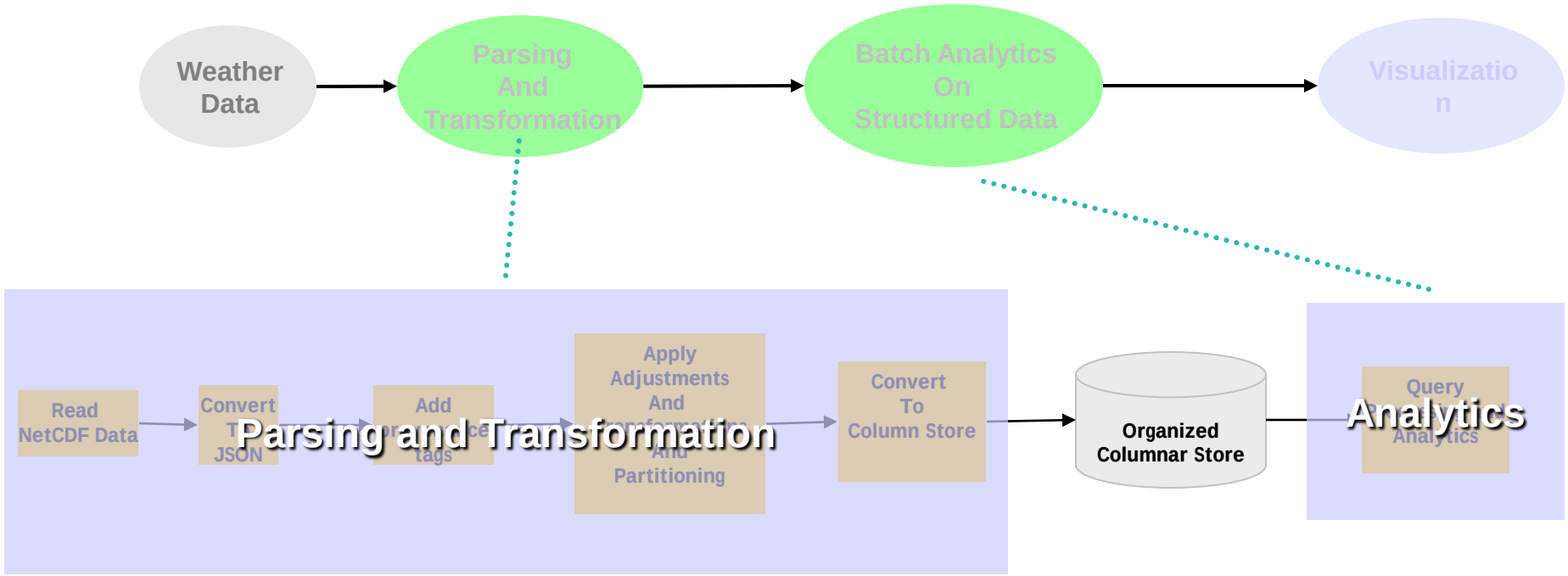
- Go beyond just indexing individual log records
- Analyze the entire data center's logs use global information for
  - Root cause analysis
  - Build advanced predictive models



Case Study: Analytics for the Renewable Energy Industry

- Gather detailed worldwide weather information from sensor measurements, modeling, and simulation
  - Attributes include wind speed, sunshine, precipitation, temperature, humidity, pressure, etc.
  - Minimum 10 year retention, possibly longer time frame: >2PB of data
- Ingest the data into an organized storage layer
- Support analytic query processing
  - Example: "Which areas had average wind speed above X mph over a 10 year period?"
  - Example: "How often does the wind speed drop below Y mph in state Z?"
- Support visualization of complex queries for domain scientists

Ingestion and Analytic Workflow for Global Weather Data



Outline

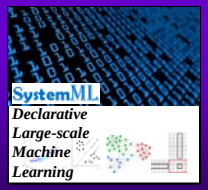
- Changing world of Enterprise Applications
- Case Studies
  - Digital Marketing Effectiveness – Media & Entertainment
  - Customer Retention and New Customer Acquisition – Retail Banking
  - System Log Analysis in a Data-Center – Cross-Industry
  - Slicing and dicing of Sensor and Simulation Data – Renewable Energy
- Quick Overview of Big Data Analytics Tools

Big Data Analytics Tools and Higher-Level Architecture

Text Analytics And Data Ingestion Engine

Entity Resolution/ Record Linkage Engine

Predictive Modeling and Machine Learning



Structured Data Processing And Aggregation

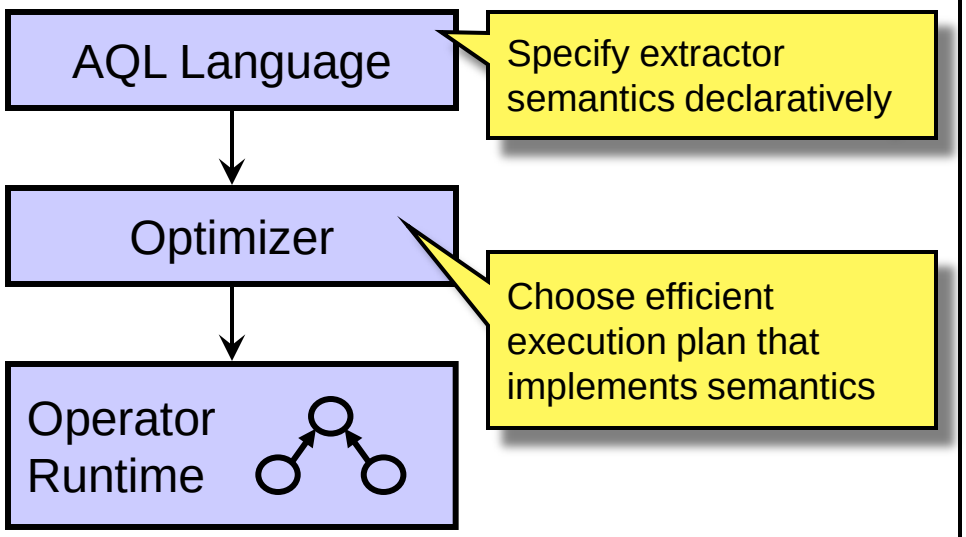
Compiler + Optimizer

Hadoop

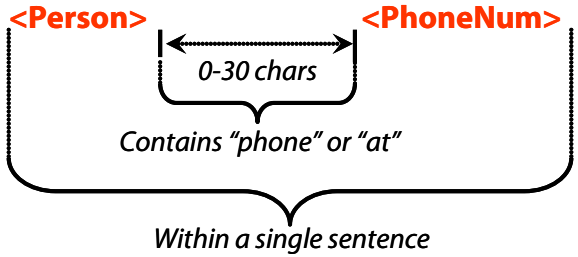
Streams

Text Analytics Engine: SystemT

SystemT Architecture



Example AQL Extractor



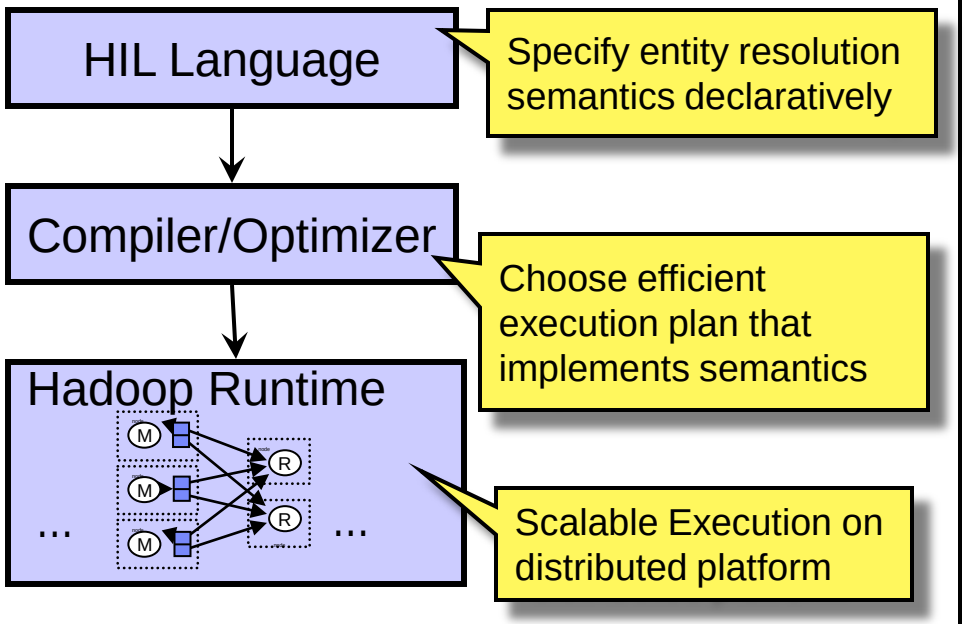

```
create view PersonPhone as
select P.name as person, N.number as phone
from Person P, PhoneNumber N, Sentence S
where
    Follows(P.name, N.number, 0, 30)
    and Contains(S.sentence, P.name)
    and Contains(S.sentence, N.number)
    and ContainsRegex(/\b(phone|at)\b/,
        SpanBetween(P.name, N.number));
```

Fundamental Results

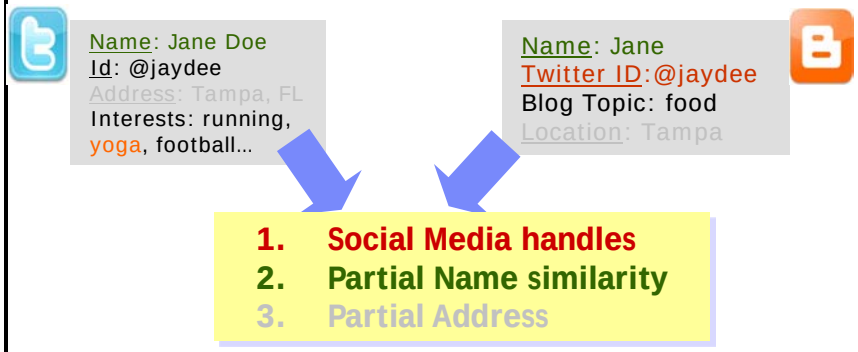
- Theorem (Expressivity): **The class of extraction tasks expressible in AQL is a strict superset of that expressible through expanded code-free CPSL grammars.**
- Theorem (Performance): **For any acyclic token-based finite state transducer *T*, there exists an operator graph *G* such that evaluating *T* and *G* has the same computational complexity.**

Entity Resolution and Record Linkage Engine

Architecture



Example Entity Resolution Rule

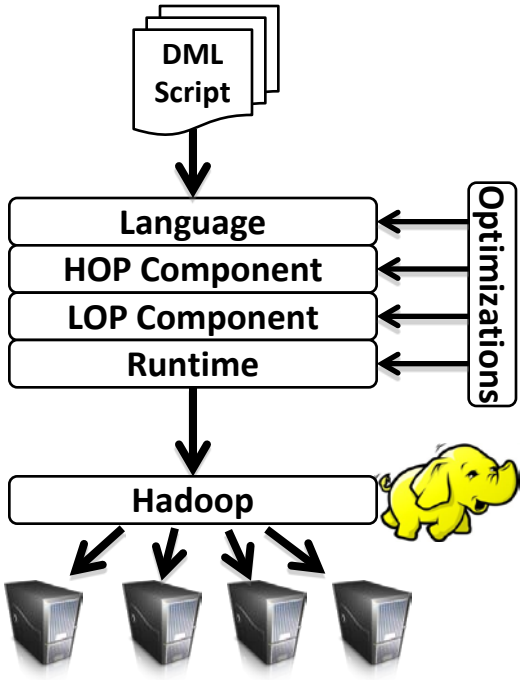



```
create "Twitter_Blogs" as
select T.Id as Tid, B.Id as Bid,
from T in $TWITTER, B in $Blogs
block T on (extractCity(Address), Id)
block B on (extractCity(Location), Twitter_ID)
where CompareUserNames(T.Name, B.Name)
    and CompareAddress(T.Address, B.Location)
    and CompareHandles(T.Id, B.Twitter_ID)
cardinality (Tid) 1:1 (Bid) ;
```

- Declarative SQL-like language over entities and relationships
- Constructs for entity definition, resolution and maintenance
- Scalable execution on MapReduce for 10's-100's millions of entities

Machine Learning Engine: SystemML

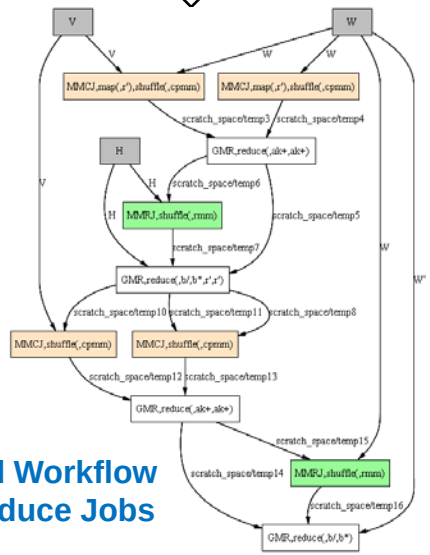
Architecture



Example DML Script for Matrix Factorization (Specification of ML algorithm)

```
1  # Input data
2  V = read("V", rows=50000000, cols=100000);
3
4  # Initial values for W and H
5  W = read("W", rows=50000000, cols=100);
6  H = read("H", rows=100, cols=100000);
7
8  max_iteration = 20;
9  i = 0;
10
11 while(i < max_iteration) {
12     H = H * ((t(W)%*%V) / (((t(W)%*%W) %*% H)+Eps));
13     W = W * ((V%*%t(H)) / ((W %*% (H%*%t(H)))+Eps));
14     i = i + 1;
15 }
16
17 # Output matrices
18 write(W, "W.out");
19 write(H, "H.out");
20
```

SystemML compilation of DML scripts

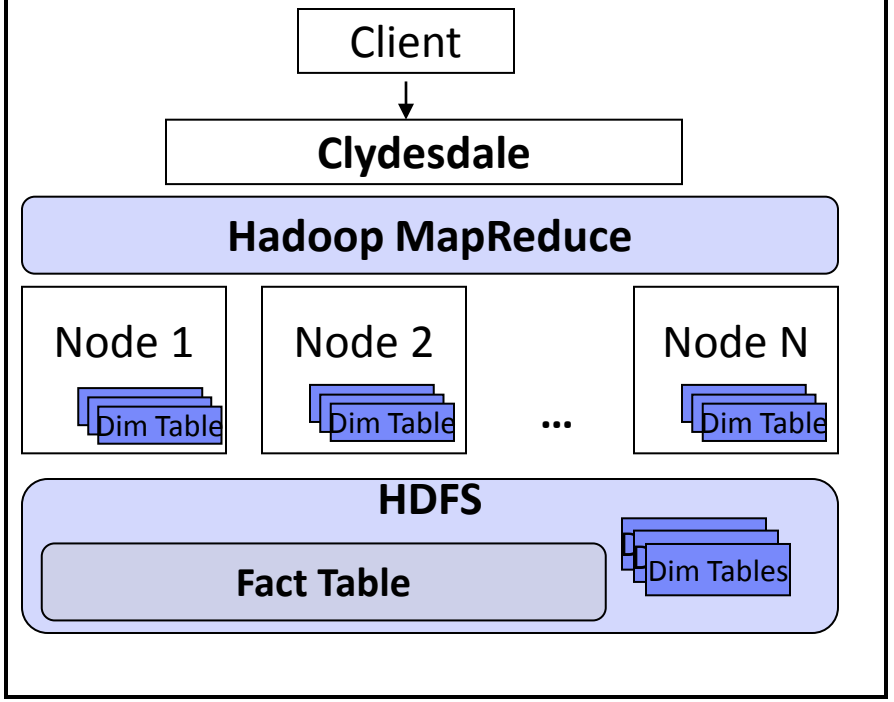


Optimized Workflow of MapReduce Jobs

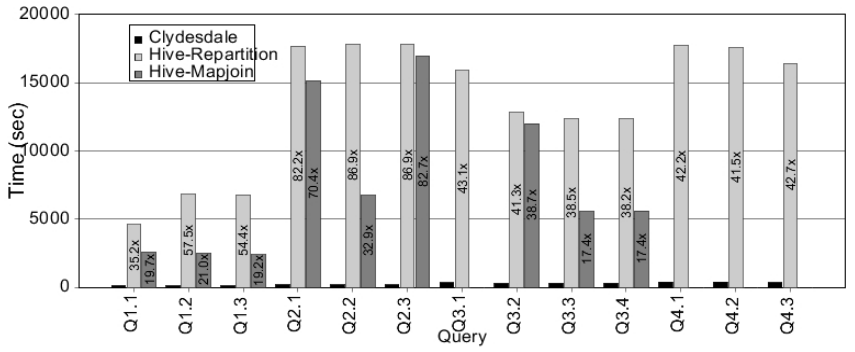
- Declarative Language with R-like syntax to ML algorithms and CV and Ensemble Learning
- DML scripts compile into efficient low-level execution plans on MapReduce
- Scales to large data volumes and massive clusters

Clydesdale: Query Processing Engine for Structured Data

Architecture



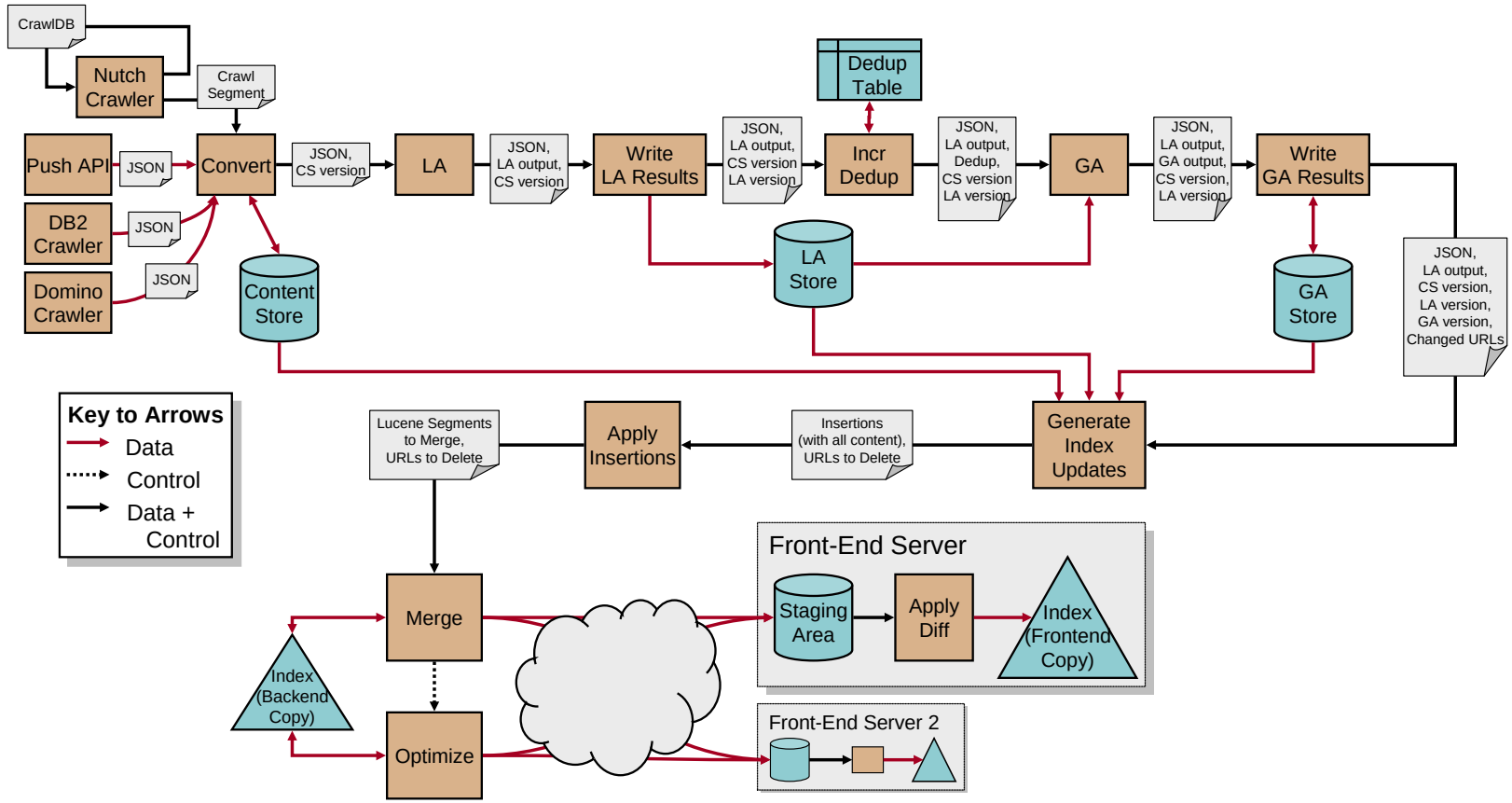
Performance Advantage over Hive (Star Schema Benchmark)



- Runtime for efficient processing of structured data
- Exploits unmodified Hadoop; orders of magnitude performance benefits over Hive
- Integrates with Jaql and other high level languages

Enterprise Knowledge Management and Search

- Generic search solution, *customizable & maintainable* in many domains
  - Simple customization with reasonable effort
  - Ongoing search-quality management



Enterprise Search

- Generic search solution, *customizable & maintainable* in many domains
  - Simple customization with reasonable effort
  - Ongoing search-quality management
- Philosophy: *programmable search*