Global Big Data Conference

www.globalbigdataconference.com

Twitter : @bigdataconf

Global Big Data Conference

Big Data Processing Beyond Map Reduce

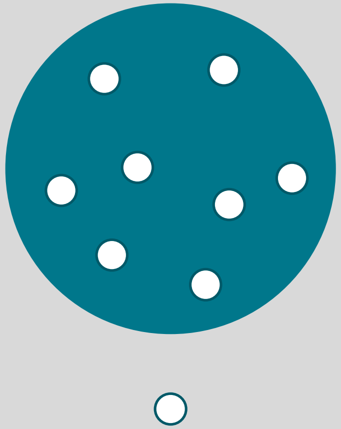
Dr. Flavio Villanustre, VP Technology
October 7, 2016

HPCC SYSTEMS®

LexisNexis® RISK SOLUTIONS

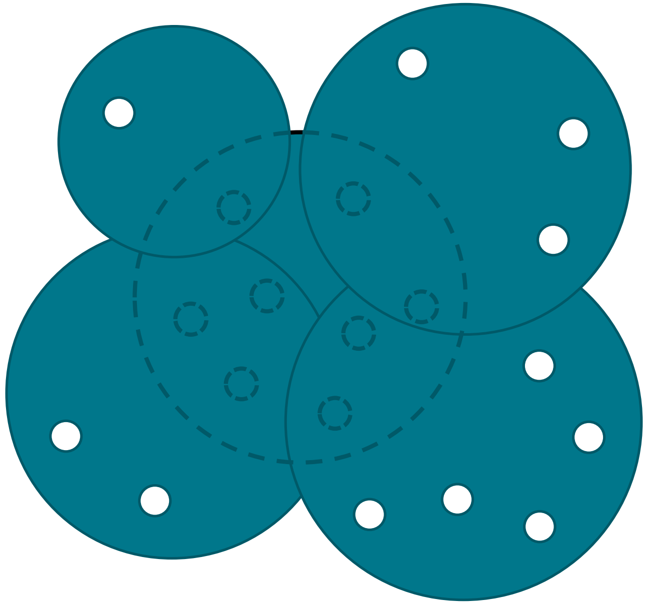
The Data Centric Approach

A single source of data is insufficient to overcome inaccuracies in the data



The holes are inaccuracies found in the data.

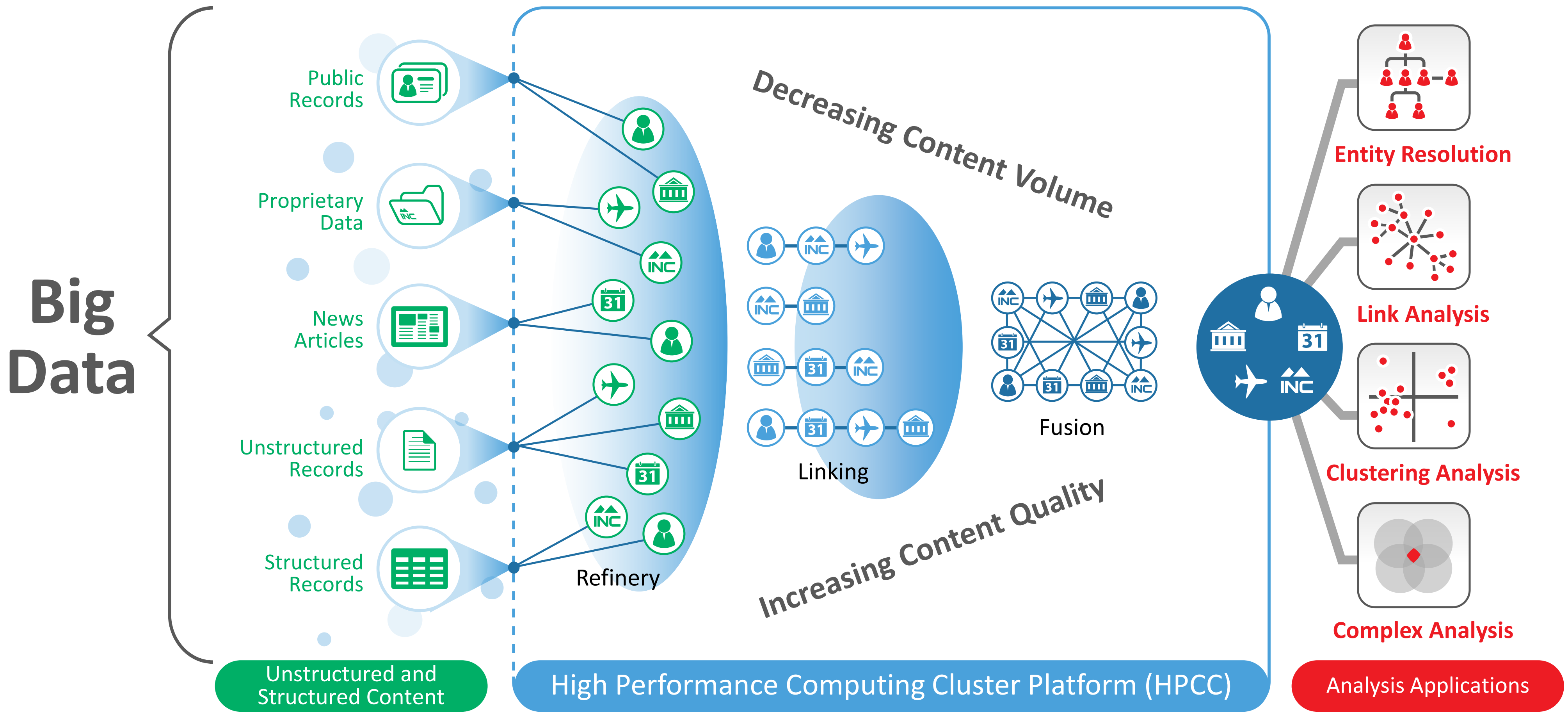
Our platform is built on the premise of absorbing data from **multiple data sources** and transforming them to a **highly intelligent social network graphs** that can be processed to non-obvious relationships.



The holes in the core data have been eliminated.

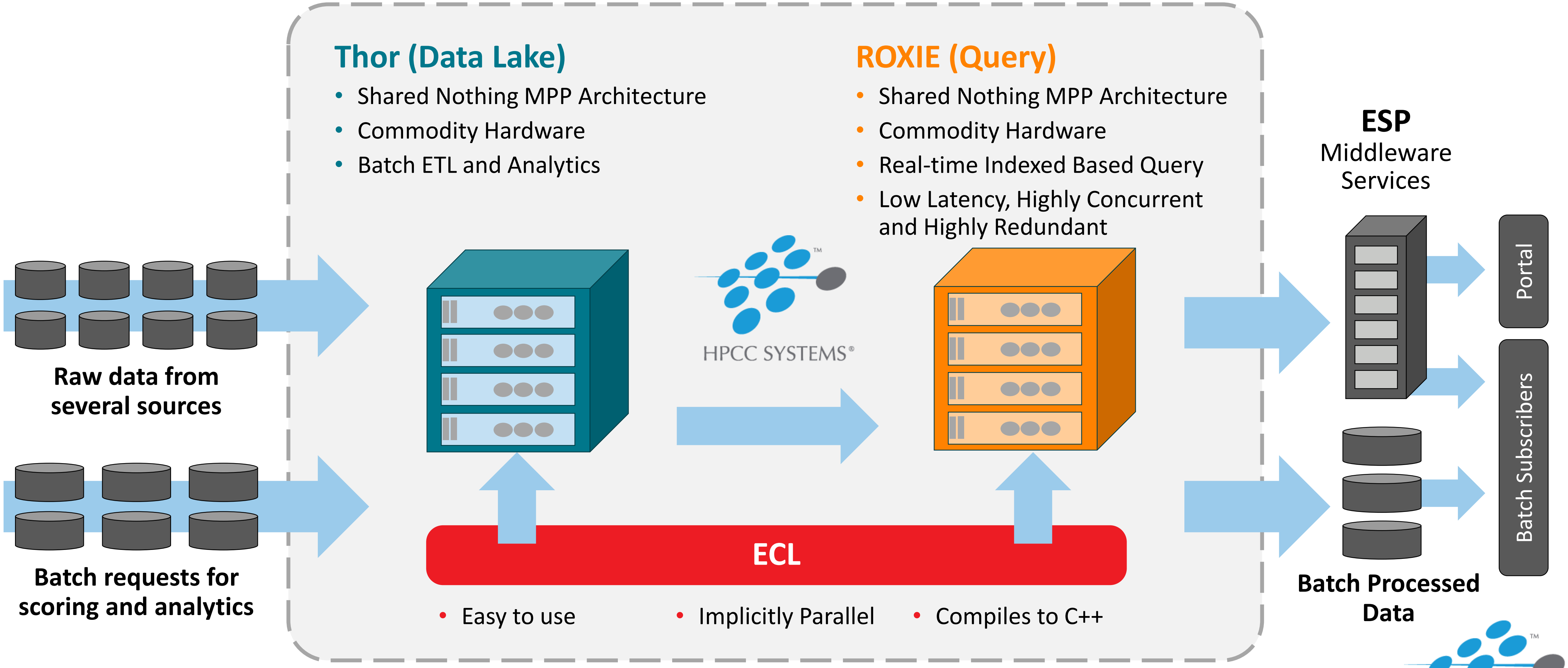
HPCC SYSTEMS®

The Big Data funnel



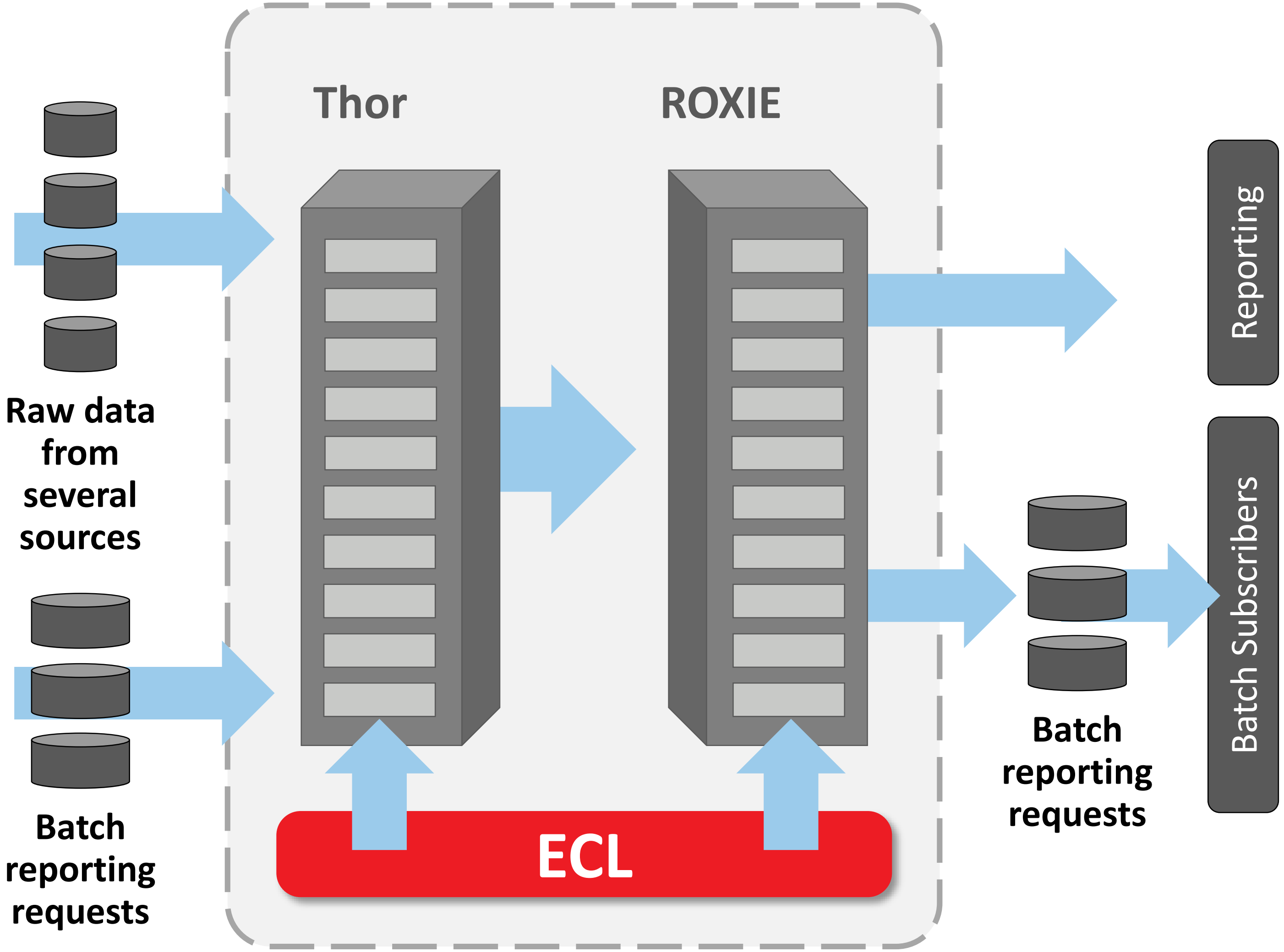
HPCC SYSTEMS®

Data Flow Oriented Big Data Platform



HPCC SYSTEMS®

ECL – The Data Flow Oriented Programming Language



- An easy to use, data-centric programming language optimized for **large-scale data management and query processing**
- **Highly efficient** — automatically distributes workload across all nodes
- **80% more efficient** than C++, Java and SQL — 1/3 reduction in programmer time to maintain/enhance existing applications
- Benchmark against SQL **(5 times more efficient)** for code generation
- Automatic parallelization and synchronization of sequential algorithms for **parallel and distributed processing**
- **Large library of built-in modules** to handle common data manipulation tasks

Declarative programming language … powerful, extensible, implicitly parallel, maintainable, complete and homogeneous

HPCC SYSTEMS®

Scalability: ECL Is Inherently — and Extremely — Parallel

FEATURES

- At the foundation level ECL works on data flows
- Imagine data is flowing through a pipe, ECL can simultaneously execute operations on different parts of the pipe
- One-to-many pipelines are constructed based on achievable parallelism and available resources
- Laziness ensures that execution is only done as needed

Original somedata dataset

01	joe chambers
02	john smith
….	….
1001	alex shaw
1002	peter king
….	….

Data Partitioned Parallel Processing

```
ds := somedata;
transformed := TRANSFORM(ds, UPPER(NAME));
```

PIPELINE 1 – **NODE 1**

direction of data flow

01 JOE CHAMBERS	02 JOHN SMITH
TRANSFORM	**Next Operation…**

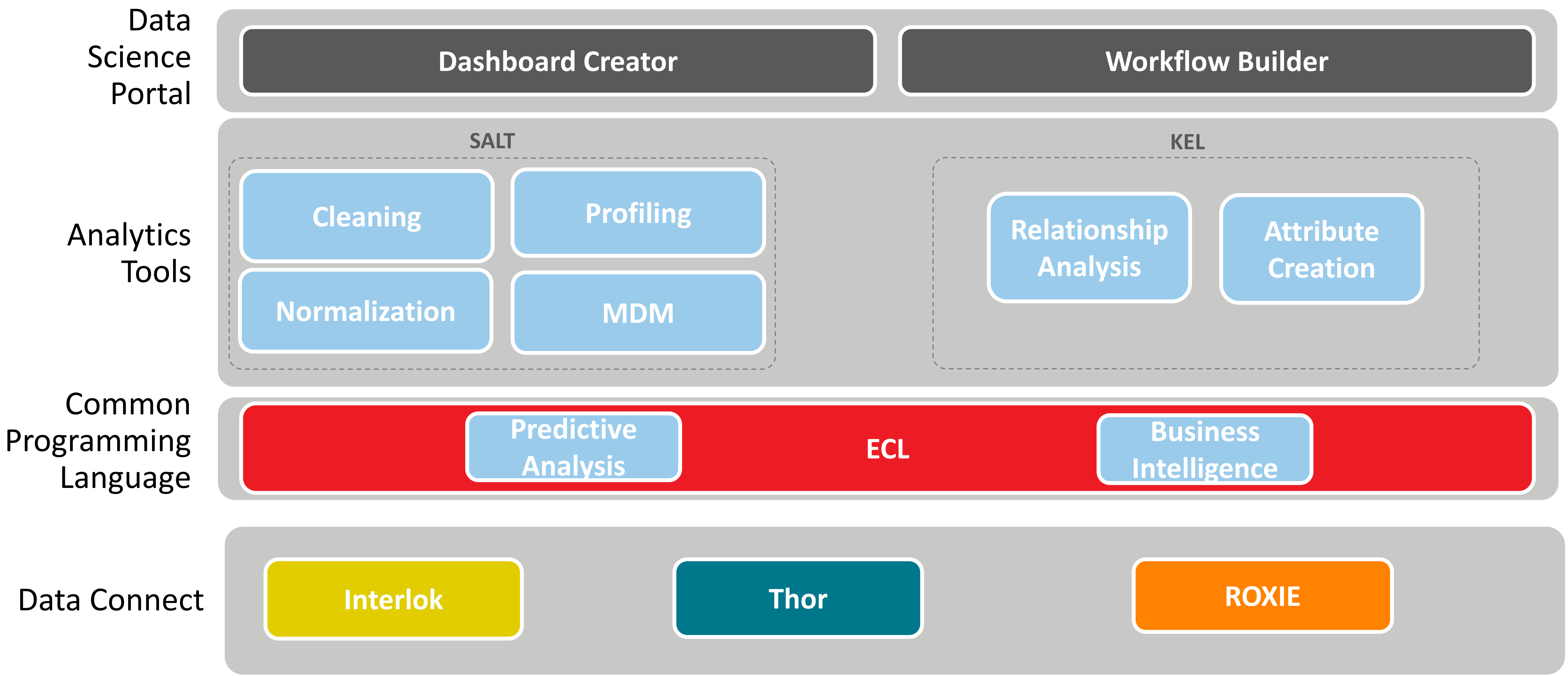
PIPELINE CC – **NODE 2**

direction of data flow

1002 PETER KING	1001 ALEX SHAW
TRANSFORM	**Next Operation…**

HPCC SYSTEMS®

The HPCC stack (STRIKE)



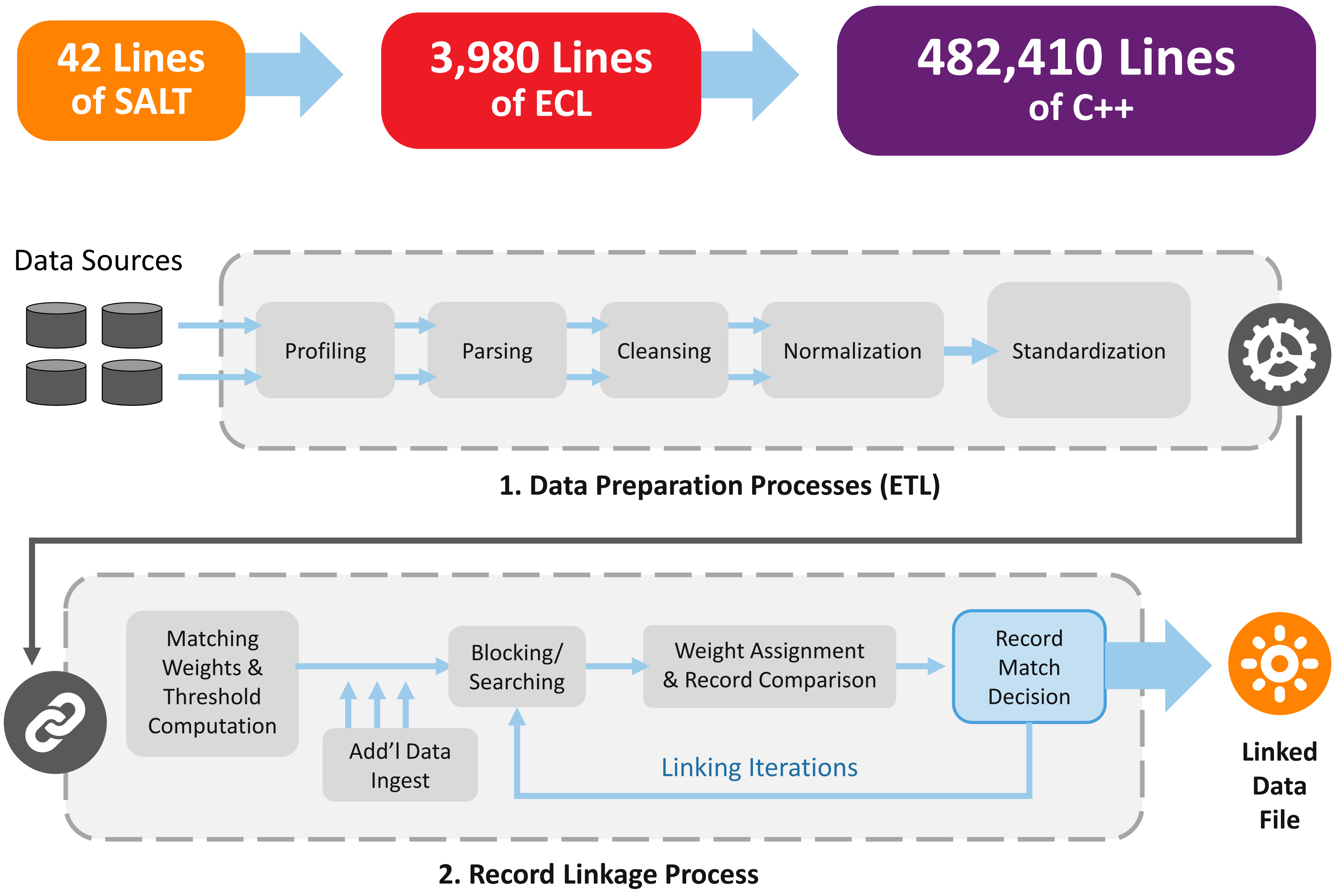
Master Data Management with SALT

From disparate data, to clustering, to showing relationships

HPCC SYSTEMS®

SALT Enables Content Disambiguation to Increase Productivity

- The acronym stands for "Scalable Automated Linking Technology"
- Entity disambiguation using Inference Techniques
- Templates based ECL code generator
- Provides for automated data profiling, parsing, cleansing, normalization and standardization
- Sophisticated specificity and relatives based linking and clustering



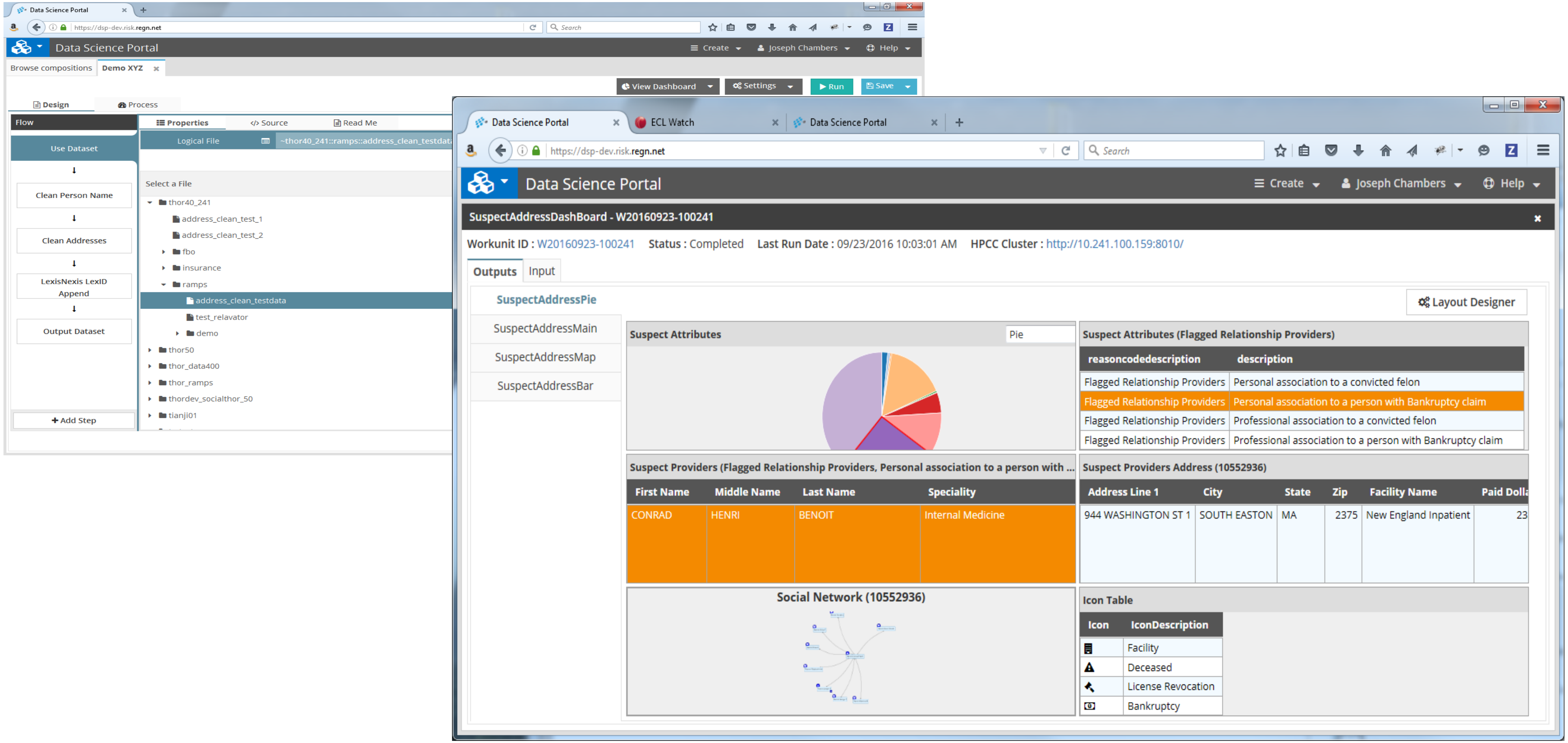
Relationship Analysis With KEL

KEL — an abstraction for network/graph processing

- Declarative model: describe what things are, rather than how to execute
- High level: vertices and edges are first class citizens
- A single model to describe graphs and queries
- Leverages Thor for heavy lifting and ROXIE for real-time analytics
- Compiles into ECL (and ECL compiles into C++, which compiles into assembler)

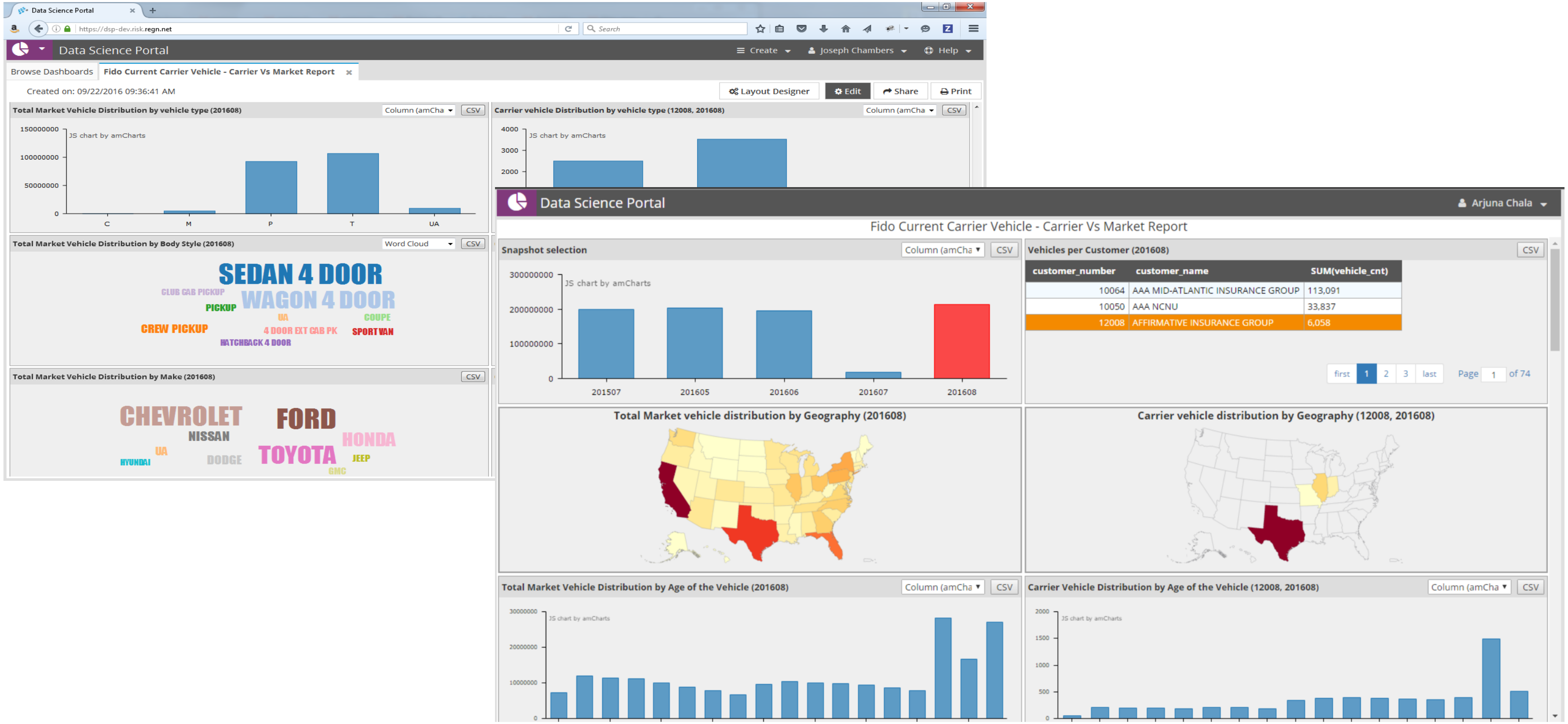
HPCC SYSTEMS®

The Data Science Portal



HPCC SYSTEMS®

Dashboards



HPCC SYSTEMS®

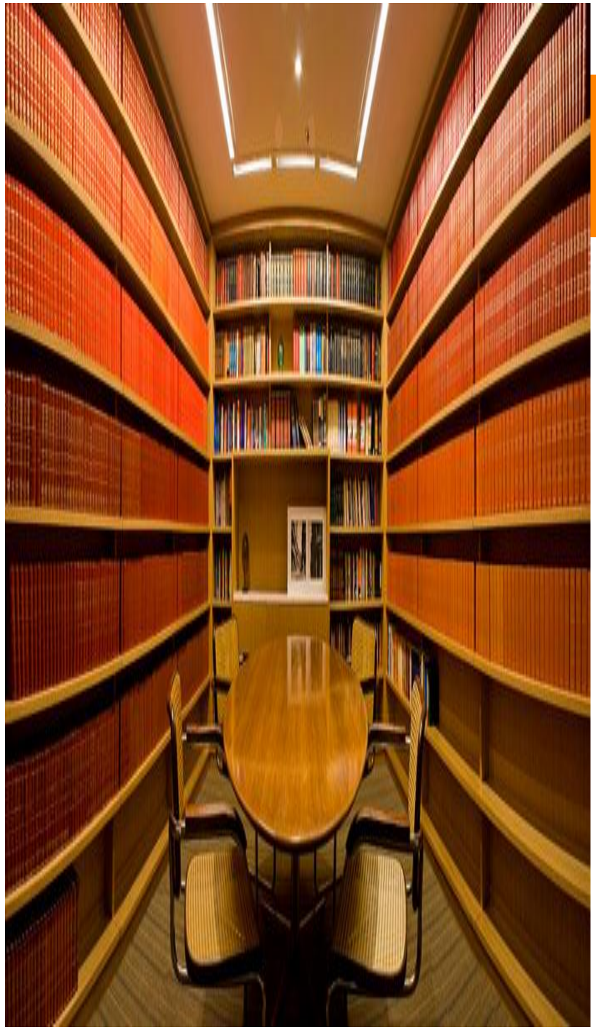
Selected use cases

LexisNexis Legal & Professional

THE CHALLENGE

- Fast insight into case law
- 100+ million documents
- Entity identification and resolution
- Document and topic classification
- Near real-time feedback

LexisNexis Legal & Professional



THE SOLUTION

- Generation 2 entity recognition employs HPCC PARSE(…) function and pattern rules
    - More entities recognized
    - Faster development
    - Faster operation
- SALT based entity resolution
- Custom resolution for citation entities
    - Case law and statute reference
    - References can be anaphoric (like *infra*)
    - Parallels (same case in more than one book)
- Active learning used to extend classification

THE OUTCOME:

Two enormous benefits:

✓ ***Huge lift on entity resolution and document classification because of SALT***

✓ ***Ahead of customers in terms of performance and maintaining currency of data because of the rapid big data processing capabilities***

HPCC SYSTEMS®

SciVal

THE CHALLENGE

- Fast insight into evidence based research data
- Scopus data covering 21,000 titles from 5,000 publishers that's updated weekly
- Data from 4,600 research institutions and 220 countries
- Real-time user specific data slicing

SciVal

THE SOLUTION

- Clean, standardize and build Analytics Queries (HPCC Thor) from 40+ TB of data on a weekly basis
- Support rapid query interface to support both ad hoc queries and canned queries (HPCC ROXIE)

THE OUTCOME:

Two enormous benefits:

✓ ***Users can customize their own visualizations, which are generated in just a few seconds***

✓ ***HPCC Systems works in 2 modes:***
1. offline crunching of huge, pre-defined requests
2. smaller calculations in real-time on customer generated data slices

HPCC SYSTEMS®

Smart Hard Hat Ecosystem

THE CHALLENGE

- 4,000 workers die and millions injured annually while working on the industrial floor
- Very high cost for maintaining safety for businesses

HPCC SYSTEMS®

Smart Hard Hat Ecosystem

THE SOLUTION

- Equip workers' hats with smart sensor technology
- Central real-time processing of (high volume) information with real-time alerting capability (HPCC Systems)
- Customizable dashboards, rules framework and data workflow frameworks (HPCC Systems)
- Predictive modeling and analytics (HPCC Systems)

THE OUTCOME:

Produced an industrial wearable that uses IoT and wireless communications systems to protect and empower industrial workers.

HPCC SYSTEMS®

Driver Behavior with Smart Telematics

THE CHALLENGE

- High cost of insurance
- High car accident rates
- Lack of tools to analyze driver behavior

HPCC SYSTEMS®

Driver Behavior with Smart Telematics

THE SOLUTION

- Telematics smart phone application
- Central system to collect (very large) data and perform analytics (HPCC Systems)
- Journey based feedback to all drivers to advice and correct behavior (HPCC Systems)
- Insurance enrollment to reduce premiums

THE OUTCOME:

- ✓ *Recommend corrections to driver behavior that would avoid accidents*
- ✓ *Reduce overall Insurance costs*
- ✓ *Correlate information from drivers data traversing the same path to create an understanding of predictable actions*
- ✓ *Examples include periods of traffic congestion, problem areas in the path and hazard detection*

HPCC SYSTEMS®

Contextual Marketing

THE CHALLENGE

Understanding an individual customer's behavior based on past actions

Technical problem

- Huge volumes of data based on observed cell phone Wi-Fi
- Apply advanced machine learning techniques

Contextual Marketing

THE SOLUTION

- Central analytics system to collect and analyze data (HPCC Systems)
- Leverage parallel algorithms to perform analytics on large quantities of data (HPCC Systems)

THE OUTCOME:

Created a platform to process any location specific telecom data that can be analyzed rapidly to gauge consumer behavior and in turn help drive context-based marketing

HPCC SYSTEMS®

Predict Passenger Volumes in Airports

THE CHALLENGE

- How to interpret 100’s millions of location points while adjusting to flight schedule changes
- Complex clustering algorithm requirements
- Understand passenger behaviors and interaction of local areas of activity

Predict Passenger Volumes in Airports

THE SOLUTION

HPCC used to solve Big Data challenges

- Raw data to refined data
- Clustering analysis
- Forecasting

THE OUTCOME:

Better passenger experience and better airport planning

Resources

- Portal: http://hpccsystems.com
- ECL Language Reference: https://hpccsystems.com/ecl-language-reference
- SALT: https://hpccsystems.com/enterprise-services/purchase-required-modules/SALT
- KEL: https://hpccsystems.com/download/free-modules/kel-lite
- Machine Learning: http://hpccsystems.com/ml
- Online Training: http://learn.lexisnexis.com/hpcc
- HPCC Systems Blog: http://hpccsystems.com/blog
- HPCC Systems Wiki & Red Book: https://wiki.hpccsystems.com
- Our GitHub portal: https://github.com/hpcc-systems
- Community Forums: http://hpccsystems.com/bb
- Case Studies: https://hpccsystems.com/resources/case-studies

HPCC SYSTEMS®

HPCC SYSTEMS®

Dr. Flavio Villanustre
VP Technology, LexisNexis® Risk Solutions
Flavio.Villanustre@lexisnexis.com

Questions?