An InterSystems Guide to the Data Galaxy

Benjamin De Boe
Product Manager
Analytics
Analytical Workflows & Tools

An InterSystems Guide to the Data Galaxy
Capturing & Storing Data

raw data  Capturing & Storing  curated data
Capturing & Storing Data

Goals:
• Persist data in a form that maximizes options for analytics

Tools:
• Storage: Hadoop, HDFS, NoSQL DBMS
• Database: RDBMS, Multi-model DBMS
• Solutions: Splunk, Flume

Challenges:
• Volume, Velocity and Variety of data
• Provide flexible & efficient access to data
Data Processing

raw data → Capturing & Storing → curated data
Data Processing

Goals:
• Transform raw data into analytics-ready formats

Tools:
• Frameworks: Hadoop MapReduce, Spark, Flink, ...
• ETL tools: ODI, Attunity, Talend, Informatica, SnapLogic, ...
• Databases: SQL, Stored procedures, ...

Challenges:
• Efficiency & scalability as volumes grow
• Trade-off between flexibility & usability
Exploration & Analytics

Goals:
- Exploratory analysis of raw data’s general characteristics & distribution
- Deeper analysis through predefined dimensions, measures & KPIs
- Identify opportunities for data refinement & further analysis

Tools:
- Exploration: DBMS, Hive, SparkSQL, Periscope, ...
- Analytics: Tableau, QlikView, Birst, Datameer, Oracle BI, Cognos, BusinessObjects, ...

Challenges:
- Trade-off between flexibility & ease of use through predefined data models
- Support for semi-structured and structured data
- Efficiency & scalability as volumes grow
Machine Learning

Business Analytics

Exploring

Processing

raw data  Capturing & Storing  curated data
Machine Learning

Goals:
- Identify correlations between input features and outcomes
- Define derived features and models that maximize predictive value

Tools:
- Frameworks: R, Spark ML, System ML, SciKit, ...
- Products: SAS, SPSS, BigML, KNIME, ...

Challenges:
- Abstract complexity of underlying algorithms
- Optimize complex computations on large datasets
Closing the Loop

Predicting → Machine Learning

Exploring

Business Analytics

Processing

Capturing & Storing
curated data

raw data
Closing the Loop

Goals:
• Apply predictive models to new data
• Leverage analytics output in transactional systems

Tools:
• Standard: PMML
• Tools: Zementis, SAS

Challenges:
• Efficient model scoring in transactional systems
• Avoid complexity and heterogeneity in production environment
Recurring Themes

Diverse tasks require diverse toolsets: **Open Platform**
- There is no single tool that covers the entire spectrum
- A proper combination of tools can help address the usability vs flexibility trade-off

Importance of a flexible storage layer: **Flexible Platform**
- Tool diversity should not be an excuse for siloing
- Move computations closer to the data to increase efficiency

Horizontal scalability is a must: **Scalable Platform**
- Almost every task lists volume-related challenges
InterSystems IRIS Analytics Strategy

An **Open Analytics Platform** allows partners and end-users to build solutions and services that leverage best-of-breed analytics technologies, both embedded in the platform and complementary to it.
InterSystems IRIS: Database

The InterSystems IRIS Database sits at the core of our Open Analytics Platform. It scales both up and out to meet the most demanding workloads:

- Parallel processing for **vertical scalability**
- ECP application servers and sharding for **horizontal scalability**

For SQL access, the InterSystems JDBC driver offers many optimizations to maximize throughput, all fully transparent to analytical tools connecting to the database:

- Leverage distributed data layout for ingestion
- Exploit low-level storage format for select queries
- Shared memory access
InterSystems IRIS: Analytics

**InterSystems IRIS: Analytics** comes with embedded technologies for exploring and analysing both structured and unstructured data. Being embedded in the platform, they can be easily leveraged in applications and solutions built on InterSystems IRIS.

**InterSystems IRIS: Business Intelligence (fka DeepSee)**

- Define and query multidimensional OLAP cubes
- Supports both ad-hoc analysis and embeddable dashboards

**InterSystems IRIS: Text Analytics (fka iKnow)**

- Identify concepts and their context from text through a unique bottom-up approach
- Enable knowledge professionals to explore and consume large volumes of texts
InterSystems IRIS: Apache Spark Connector

Apache Spark addresses the need for efficient, horizontally scalable data processing through its **Resilient Distributed Dataset** abstraction

- Much higher developer productivity than Hadoop’s MapReduce framework
- Lazy execution & optimized code generation enables very high throughput
- Comes with broad set of libraries, including Machine Learning, Graph, etc
- Active and broad user community

Apache Spark’s Data Source API allows database vendors to implement optimizations that leverage the underlying data store

- Spark’s Catalyst compiler will use these to push computations to the data
- Further convenience extensions are also available
InterSystems IRIS: Apache Spark Connector

InterSystems IRIS Apache Spark Connector capabilities:

• Convenient Read/Write API
  – Expose DB table as Dataset
  – Expose arbitrary SQL query as Dataset

• Pushdown strategies
  – filter(), select(), and more advanced where possible

• Leverage distributed DB data layout & parallelism
  – Biggest gains vs vanilla JDBC connector
InterSystems IRIS: Apache Spark Connector

![Diagram of InterSystems IRIS: Apache Spark Connector](image-url)
InterSystems IRIS: PMML Support

Transactional System

Model Consumer

Advanced Analytics

Model Producer
InterSystems IRIS: PMML Support

Transactional System

Model Consumer

PMML

Advanced Analytics

Model Producer
InterSystems IRIS: PMML Support

PMML standard enables best-of-breed approach for Predictive Analytics
• Leverage existing skillset in 3rd party tools for model building in lab environment
• Leverage InterSystems IRIS Data Platform for model execution in production environment
• Fits our Open Analytics Platform roadmap

Standard support in InterSystems IRIS: PMML Consumer
• Automatically generate optimized COS code based on PMML
• Leverage from other InterSystems IRIS embedded technologies, …
InterSystems IRIS: UIMA Integration

Combining different NLP tools is challenging at best
InterSystems IRIS: UIMA Integration

UIMA standard offers interoperability & scalability for NLP application developers
InterSystems IRIS: UIMA Integration

InterSystems IRIS UIMA Integration marries data platform and UIMA data processing infrastructure for increased developer productivity and a quicker path to analytics.
Conclusion

An InterSystems Guide to the Data Galaxy
Conclusions

The InterSystems IRIS Data Platform is an Open Analytics Platform, helping customers leverage the tools they know on a platform that scales:

- Through **embedded technologies** for exploration & analytics
- Through connectors for best-of-breed **third-party technologies**, allowing those tools to leverage our platform in a way that optimizes overall system performance
- Through integration with relevant **industry standards**, enabling customers to leverage models and components created in dedicated tools as part of new solutions
Related Sessions

InterSystems IRIS: Database
• What’s Lurking in your Data Lake? (Tue, 3PM)
• We Want More, Solving Scalability (Tue, 4PM; Wed, 11AM)
• Speeding up JDBC (Wed, 10:30AM, flash talk)

InterSystems IRIS: Spark Connector
• Apache Spark Experience lab (runs daily)

InterSystems IRIS: Text Analytics
• iKnow: Treating Patients with iKnow and REST (Mon, 3:30PM)
• iFind: Catching Bad Guys with iFind and REST (Tue, 4PM)
• UIMA: Annotations on the Sofa (Wed, 9:45AM, flash talk)
Please complete the survey in the event app.

Session recording and slides will be available at: http://Learning.InterSystems.com
Search for “Global Summit 2017”

Visit the Tech Exchange to learn more!
Thank you.