Celgene

The Data Catalog
The Key to Managing Data, Big and Small

April Reeve May 18 2017

April Reeve

- Thirty years doing data oriented stuff

Data Stuff

- **Data Management discipline – Data Integration, Data Governance, Data Modeling, Data Quality, Business Intelligence, Master Data Management, Data Conversion, Data Warehousing , Enterprise Content Management, Big Data Management**
- Currently **Director of Enterprise Information Strategy and Architecture** at **Celgene Corporation** in Summit, NJ
- Certifications –
  - **Certified Data Management Professional (DAMA CDMP)**
  - **Certified Data Governance and Stewardship Professional (ICCP DGSP)**
  - **Certified Business Intelligence Professional (TDWI CBIP)**
  - Certified in Enterprise Governance of IT (ISACA CEGIT)
  - Certified Information Systems Auditor (ISACA CISA)
- Masters degree in Financial Management (predictive modeling, risk management, derivatives, corporate finance) from Erasmus University, Rotterdam
- **Book “Managing Data in Motion – Data Integration Best Practice Techniques and Technologies”**
- **New chapters - Data Management Body of Knowledge (DMBoK) release 2 – Data Integration and Big Data**

Agenda

Why is the data catalog now a critical data management component?

Big Data

Enterprise Data

Creating the data inventory and the data catalog

Types of Metadata

- Business metadata - provides the meaning of the data. It helps define terms in every-day language, without regard to technical implementation.
  - Examples: business rules, stewardship, business definitions, auditing terminology, glossaries, algorithms, and lineage using business language • Audience: business users
- Technical metadata – provides information on the format and structure of the data as needed by computer systems.
  - Examples: definition of source and target systems, their table and fields structures and attributes, documentation for auditing derivations and dependencies • Audience: specific tool users (BI, ETL, profiling, modeling)
- Operational metadata - refers to the metadata generated and captured when a process executes. It allows administrators to manage the system and ensure things are running smoothly. Includes audit trail information – what was updated or accessed when and by whom.
  - Examples: information about application runs, including their frequency, record counts, component-by-component analysis, and other statistics for auditing purposes • Audience: operations, management and business users

Why is the Data Catalog an important component?

Why now?

Big Data

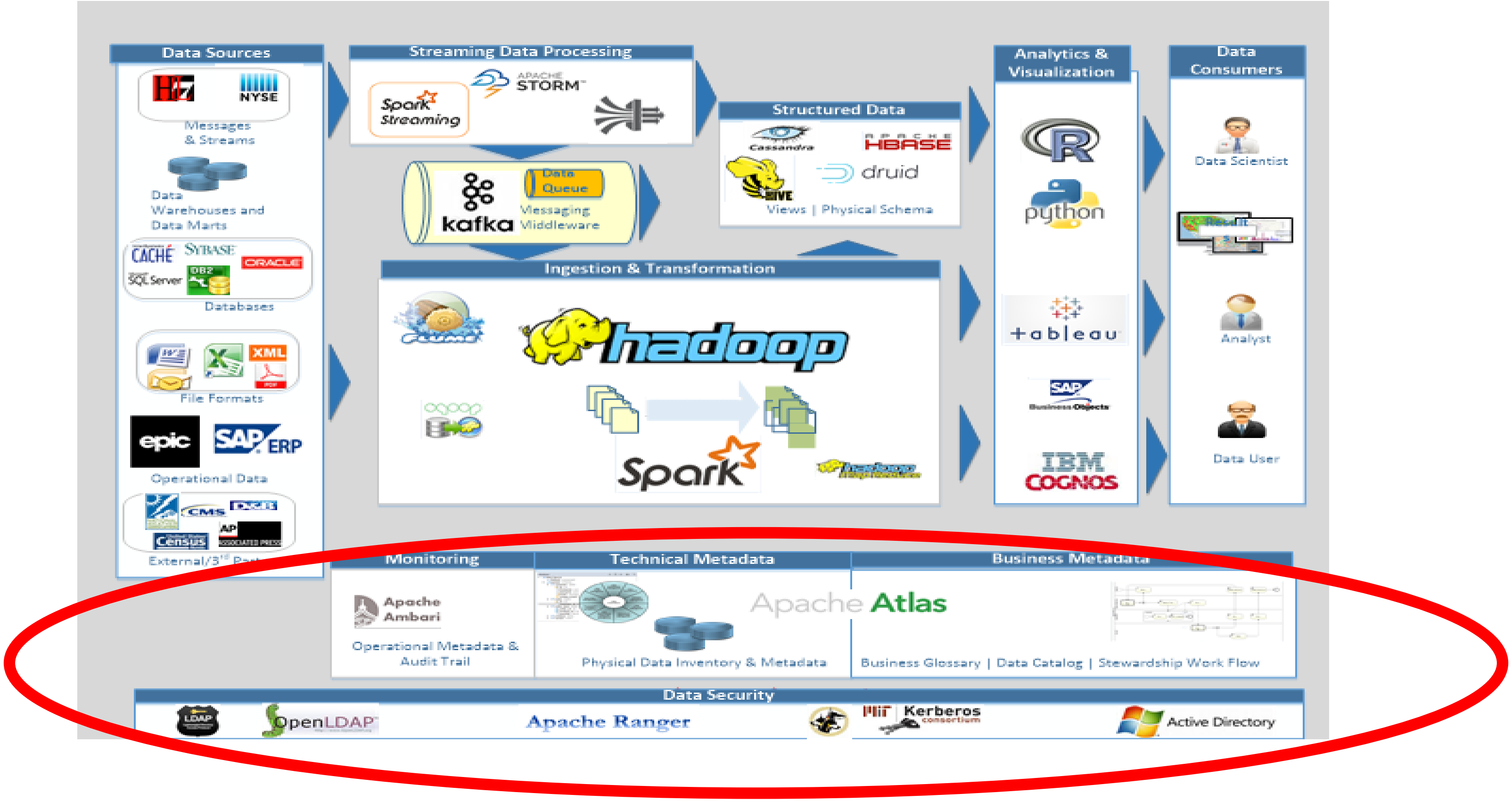
Big Data Technology Limitations

- All of the analyst organizations, including Gartner Group and Forester, have predicted that the lack of built in metadata repositories in Big Data technology (such as Hadoop) will be one of the biggest risks of project success for Big Data projects.
- Few Big Data solution architects have the expertise to understand what functional capabilities they are missing: data access security solutions (usually by role and asset), audit trails of data update and access (operational metadata), inventory of data assets (technical and business metadata).

“Through 2018, 80% of data lakes will not include effective metadata management capabilities, making them inefficient”

- Gartner Group, September 2016

Data Lake Architecture



Data Sources

Messages & Streams

Data Warehouses and Data Marts

Databases

File Formats

Operational Data

External/3rd Party

Streaming Data Processing

Data Queue

Messaging Middleware

Structured Data

Views | Physical Schema

Ingestion & Transformation

Analytics & Visualization

Data Consumers

Data Scientist

Analyst

Data User

Monitoring

Operational Metadata & Audit Trail

Technical Metadata

Physical Data Inventory & Metadata

Business Metadata

Business Glossary | Data Catalog | Stewardship Work Flow

Data Security

Data Governance of Big Data

- Traditional Data Governance
  - Data Issue Tracking
  - Business Metadata / Business Glossary
  - Technical Metadata
  - Data Quality Improvement
  - Metrics Tracking & Reporting
  - Data Profiling & Monitoring

- Big Data Governance
  - All those plus:
  - Where capture metadata?
    - Business meaning
    - Technical structure and format
    - Audit trail
  - Volume - Can metadata capture be automated?
  - Variety / velocity - Automatically consolidate metadata from various repositories / data types?
  - Variety - How profile and monitor data?
  - Do data stewards need new technical skills?

Enterprise Data

Providing Data as a Service

- IT organizations need major improvements in their data management processes – especially in the basic process of providing data to users (people and systems)
- In order to manage something, the first step is to inventory it. How many organizations have an inventory of their data?
- There needs to be processes for data consumers to identify available data, request access, receive appropriate review and approval, and have data provisioned, without taking huge amounts of time and IT effort

Data Catalogs & Data Services

- Data Catalogs – menu of available data
  - Data Lake / Big Data Implementations must maintain an Inventory – briefly, what is each file & who is responsible (business metadata) and, if possible, what is the technical layout
  - A Data Catalog is a menu of what data is available from which a User selects and, if access approved, data is provisioned
- Data Services – business functions to be improved
  - Request access to data
  - Request additional data added
  - Request a report, analysis (root cause, data quality assessment, exploratory), a predictive model
  - Request operational system changes based on a prediction

Creating the data inventory and the data catalog

Integrating metadata into a central data inventory

- Every technology development tool has a metadata repository (except Hadoop?)
- Integrating the metadata data together should be as automated as possible
- Sourcing business metadata and linking to technical metadata may be a challenge
  - Data models
  - Data dictionaries
  - Business glossaries
- Technical metadata must be updated on a periodic, automatic basis but business metadata may not be updated as frequently
- This is as challenging as any data integration project

Data Stewardship

- Who is the responsible business person(s) for each set of data?
  - Subject Matter Expert to answer questions
  - Business owner to approve access
- Who is the responsible technical person(s) for each set of data?
  - Database administrator
  - Application manager
  - Access manager to provide access when receive business approval
- Automatic monitoring for missing information and sending alerts / requests to appropriate Data Stewards
- Data Steward front end

Using the data catalog to request access

- The integrated metadata is an inventory but not a catalog
- Need a business front end to:
  - Provide a list of the data (and data services)
  - Provide a way to request access to data
  - Kicks off an automated work flow to request access from the “data owner”
  - Kicks off an automated work flow to grant access by the appropriate technical person
    - Provide access to the data in place?
    - Provide a copy of the dataset?
    - Schedule periodic copy/update to a data sandbox?

April Reeve

areeve@Celgene.com

@Datagrrl on Twitter

Book - "Managing Data in Motion – Data Integration Best Practice Techniques and Technologies"