IBM InfoSphere Datastage and Hadoop

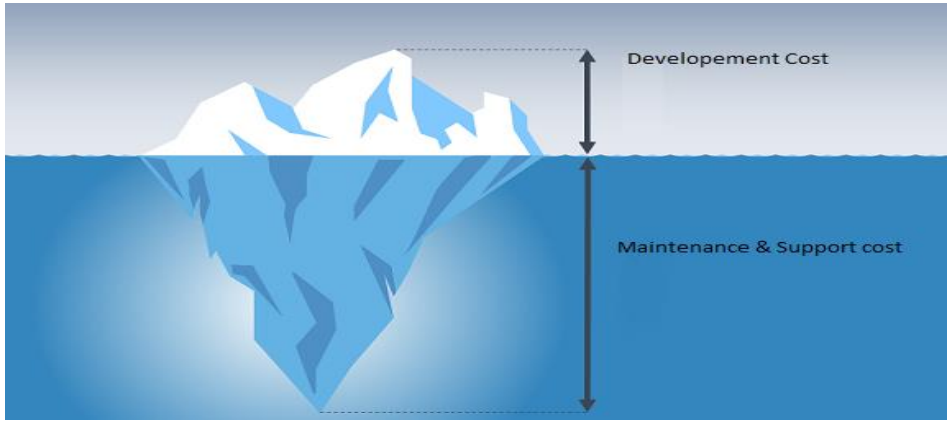
- Two Best-of-Breed Solutions Together

Abstract

In the era of Big Data, the amount of data organizations deal with, has grown exponentially changing the traditional ways of data management. Much of the early development on Hadoop has been done through scripting and heavy coding including JAVA, MR, Pig, Sqoop etc. which increases migration and operational cost.

The practice of hand-coding brings back the argument from early data warehousing days on the usage of hand-coded jobs vs ETL tools. For the same reasons the ETL tools were chosen as the go to option back then, data integration tools like InfoSphere DataStage is becoming the preferred option in building Time to Market Data Lakes and other Big Data solutions without compromising on Hadoop native features and providing 10x scale parallel processing by virtue of DataStage MPP architecture.

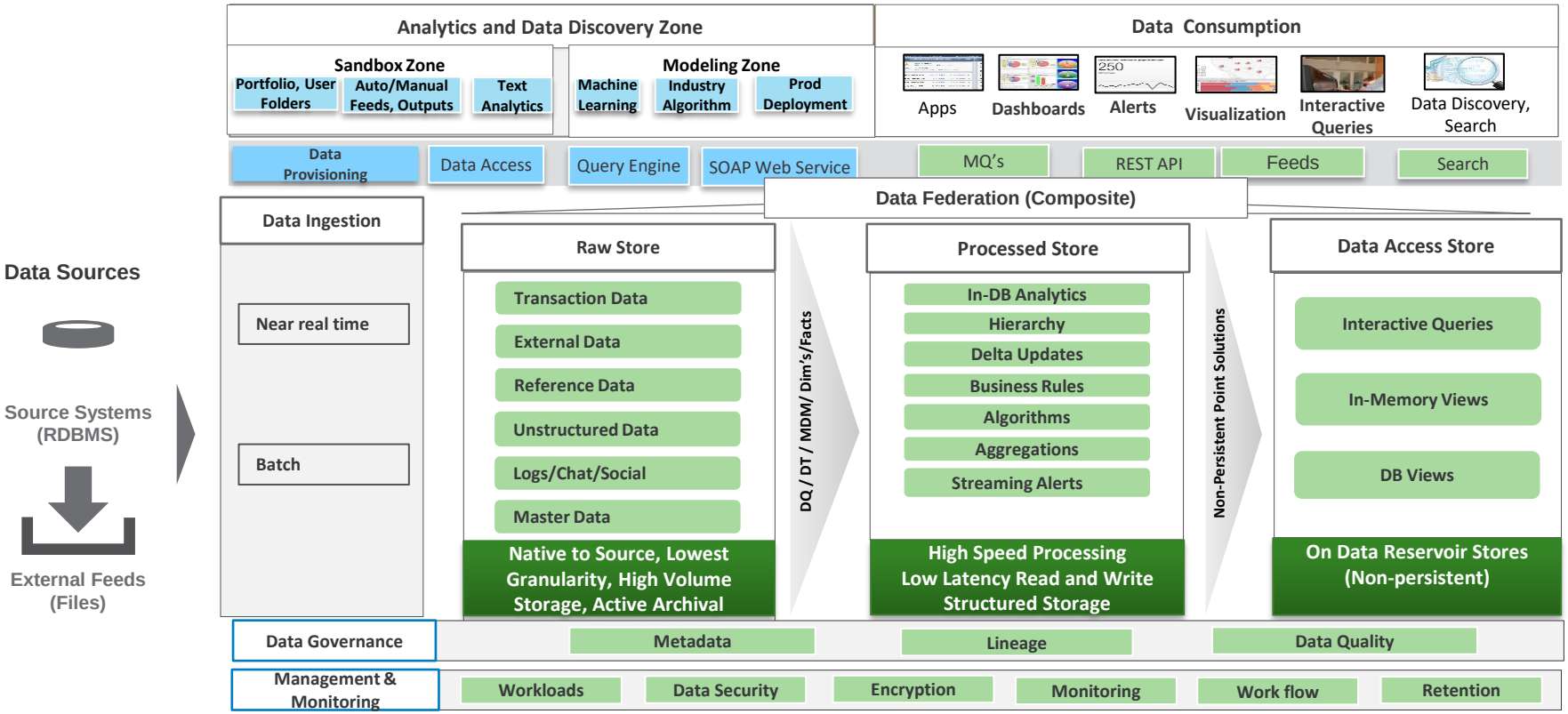
Hand-coding vs. Adoption of ETL



Why should I beware of Hand coding Iceberg ?

- Time to Market
- Re-usability
- Security & scalability
- Build once and run anywhere
- Easy Adoption & Integration
- Reduction in cost
- Data Governance & lineage
- Enhanced DQ & visualization

Datalake – ‘The Visionary’



Our journey through Datalake using Datastage …

Why Datastage for Hadoop ?

“Hadoop is not a data integration solution”

“Although many Hadoop projects perform ETL work streams, Hadoop lacks the necessary key features of commercial data integration tools”

“Metadata management, DQ, data lineage and data integration, among other things, are crucial prerequisites for a successful data lake. They cannot be afterthoughts.”

- Gartner Group Research Note , January 2013

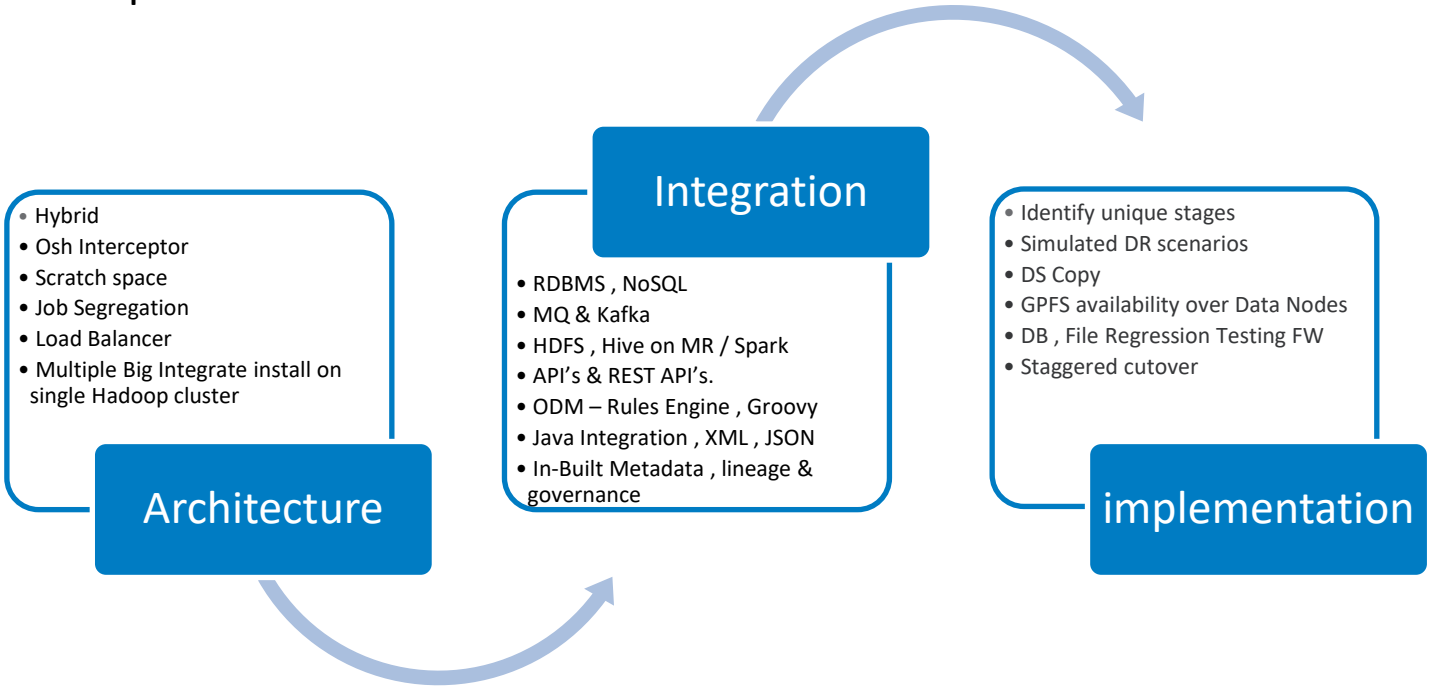
Our journey through Datalake using Datastage …

- Offers Extensive Data Lake feature with multiple source Integration
- Port existing DataStage jobs to run on Hadoop
- Reduction of storage cost
- Push down architecture which is in-line with Hadoop architecture
- IIS has better solution of Data cataloging and Unified Governance.
- UI based coding to reduce the time & effort
- No proven ETL process which integrates well with Hadoop
- Process can run in Native PxEngine or in Spark. Co-exist with other Hadoop tech-stacks. Massively scalable.
- Clients will spend Money in re-shaping their process in data lake for better Insights & ML rather spending money in hand coding
- Fault tolerant , Flexible adoption of other tech stacks.

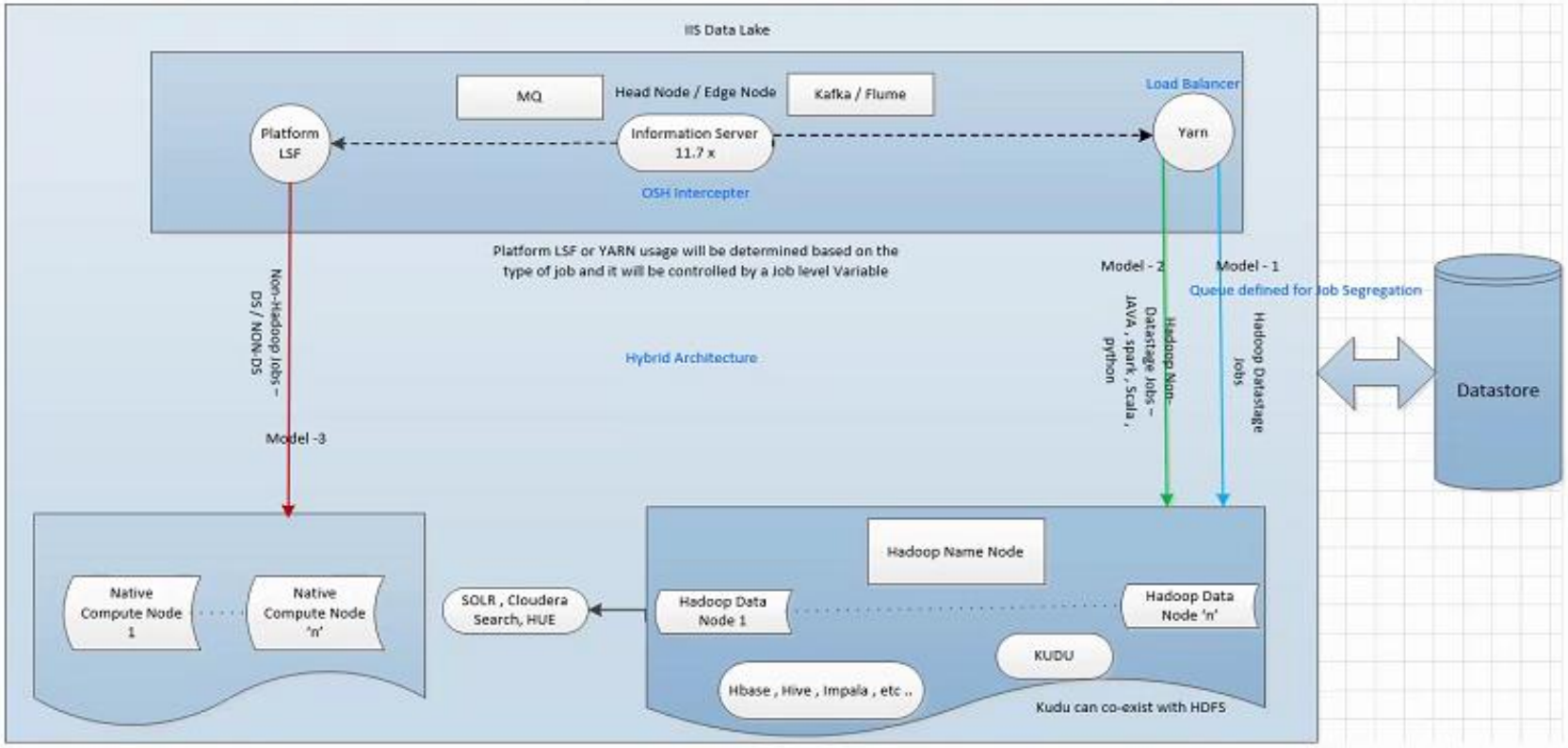
Best-of-Breeds Together

Our journey through Datalake using Datastage …

- For a successful migration/implementation due diligence need to be done on all different aspects



Architecture :



Integration

Database

Big-Data

Streaming

Implementation

Idea’s are easy but implementation is hard

- Identify unique stages
- Simulated DR scenarios
- DS Copy
- GPFS availability over Data Nodes
- Database & File Regression Testing FW
- Staggered cutover

Performance tuning

Optimization is not a science, it's an Art

- Achieve maximum concurrency
- Optimal usage of container vs Vcore
- Application Master pre-emption
- Application Master Related APT variables
- YARN configuration parameters
- Data Locality

Appendix

Our journey through Datalake using Datastage

Why Datastage ?

The existing 10's and 1000's of jobs coded in Datastage can be ported as-is in to Hadoop and built Data Lake without any increase in cost.

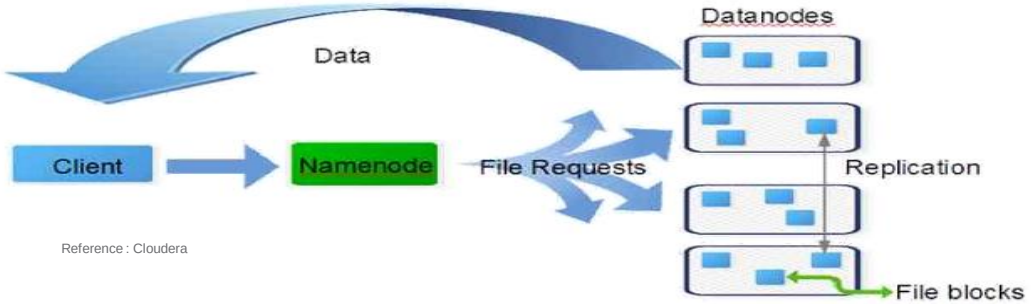
The Native clusters will be replaced with Hadoop Data nodes to reduce the storage cost when compares to SAN which will be used for one time Datastage version upgrade.

The PX engine and the Spark (in-flight development) observed as the best processing engines for Data Lake which supports Ingestion , processing and provisioning with better Data cataloging and Unified Governance.

Reference : IBM

Why Hadoop ?

- Open source , many free version available
- Cost effective , runs on commodity hardware
- Massively scalable and distributed for processing large volumes
- Fault tolerant, losing nodes doesn’t mean losing your data or failed jobs
- Flexible, many tools built on top for analytics and data movement
- Moves processes near to data
- Scalability



Reference : Cloudera

Our journey through Datalake using Datastage …

Best-of-Breeds Together

- Port Existing Datastage Jobs
- Reduction of storage cost
- Ingestion , Transform & Provisioning
- Data cataloging and Unified Governance.
- UI development

- Open source
- Cost effective , runs on commodity hardware
- Massively scalable
- Fault tolerant
- Flexible, many tools built on top for analytics and data movement