July 20, 2021

Scaling S3 Parquet Scanning Performance using EMR

Presenter, Designation

Informatica™

Housekeeping Tips

- Today’s Webinar is scheduled for 1 hour
- The session will include a webcast and then your questions will be answered live at the end of the presentation
- All dial-in participants will be muted to enable the speakers to present without interruption
- Questions can be submitted to “All Panelists" via the Q&A option and we will respond at the end of the presentation
- The webinar is being recorded and will be available on our INFASupport YouTube channel and Success Portal - where you can download the slide deck for the presentation. The link to the recording will be emailed as well.
- Please take time to complete the post-webinar survey and provide your feedback and suggestions for upcoming topics.

Feature Rich Success Portal

- Bootstrap trial and POC Customers
- Enriched Customer Onboarding experience
- Product Learning Paths and Weekly Expert Sessions
- Informatica Concierge
- Tailored training and content recommendations

Informatica

More Information

Success Portal

https://success.informatica.com

Communities & Support

https://network.informatica.com

Documentation

https://docs.informatica.com

University

https://www.informatica.com/in/services-and-training/informatica-university.html

Safe Harbor

The information being provided today is for informational purposes only. The development, release, and timing of any Informatica product or functionality described today remain at the sole discretion of Informatica and should not be relied upon in making a purchasing decision.

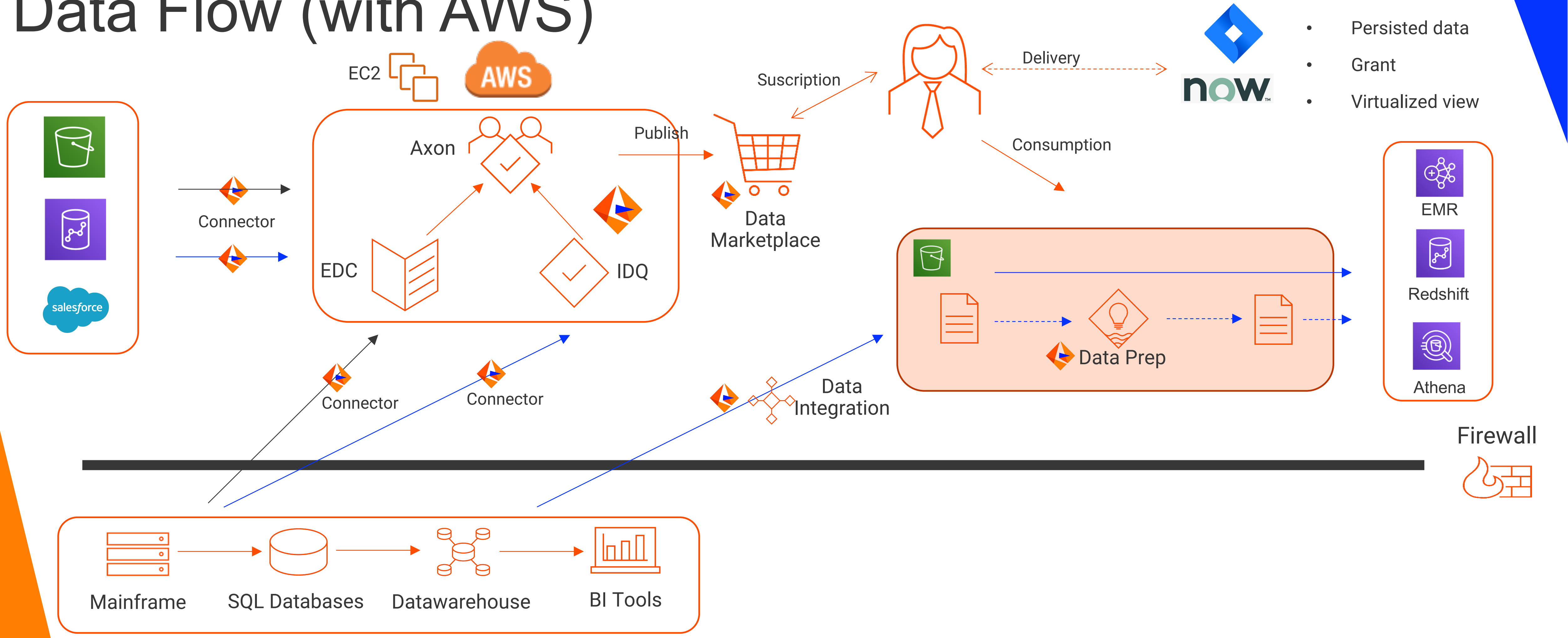
Statements made today are based on currently available information, which is subject to change. Such statements should not be relied upon as a representation, warranty or commitment to deliver specific products or functionality in the future.

Agenda

- UseCases
  - EDC: Scan Data Lake in S3, Domain Tagging
  - DEQ: Data Quality Profiling, Mapping and Scorecarding
- Data Source Preparation
- Maintenance
- Configuration
- Infrastructure

UseCases

Data Flow (with AWS)

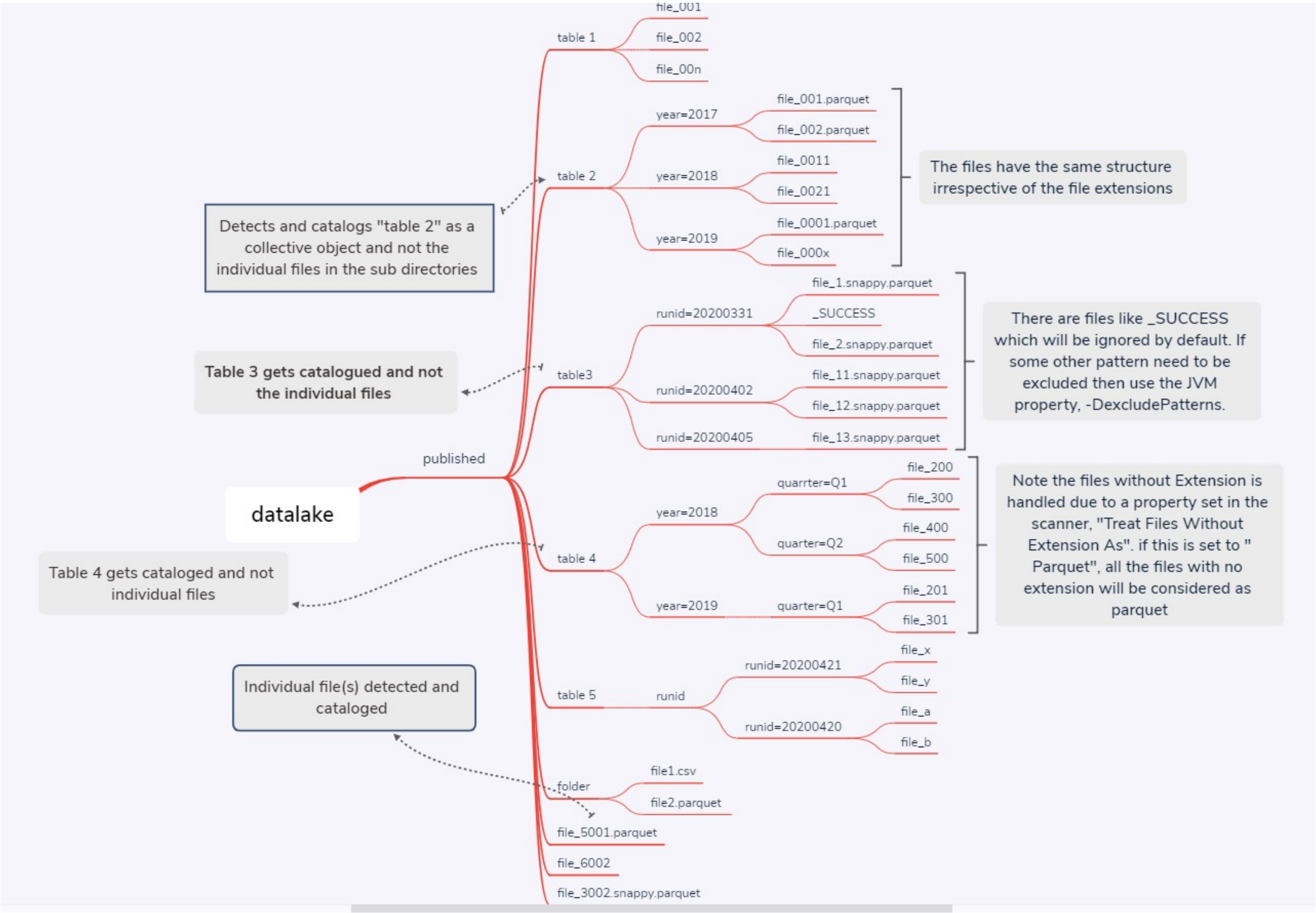


- Persisted data
- Grant
- Virtualized view

→ Metadata & Data

→ Data

EDC: Partition Detection



Detects and catalogs "table 2" as a collective object and not the individual files in the sub directories

The files have the same structure irrespective of the file extensions

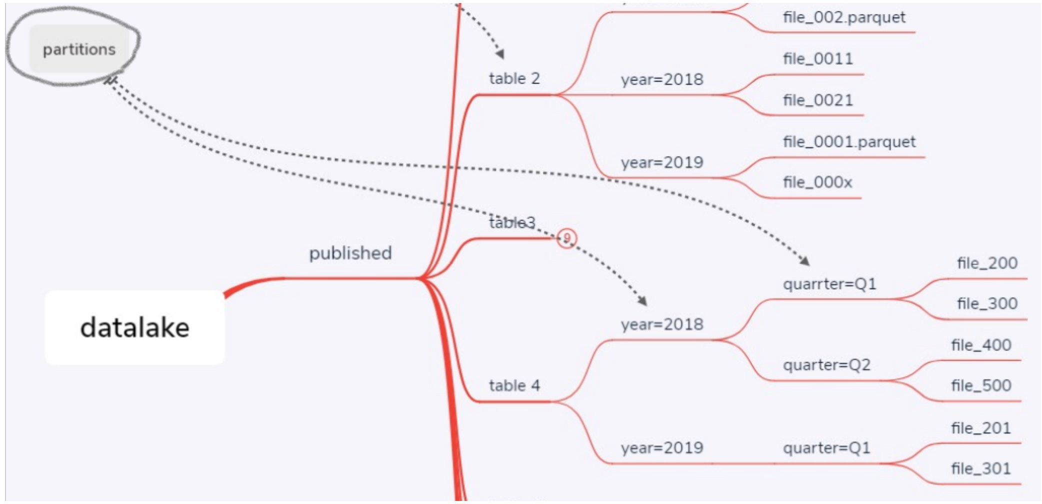
Table 3 gets catalogued and not the individual files

There are files like _SUCCESS which will be ignored by default. If some other pattern need to be excluded then use the JVM property, -DexcludePatterns.

Table 4 gets cataloged and not individual files

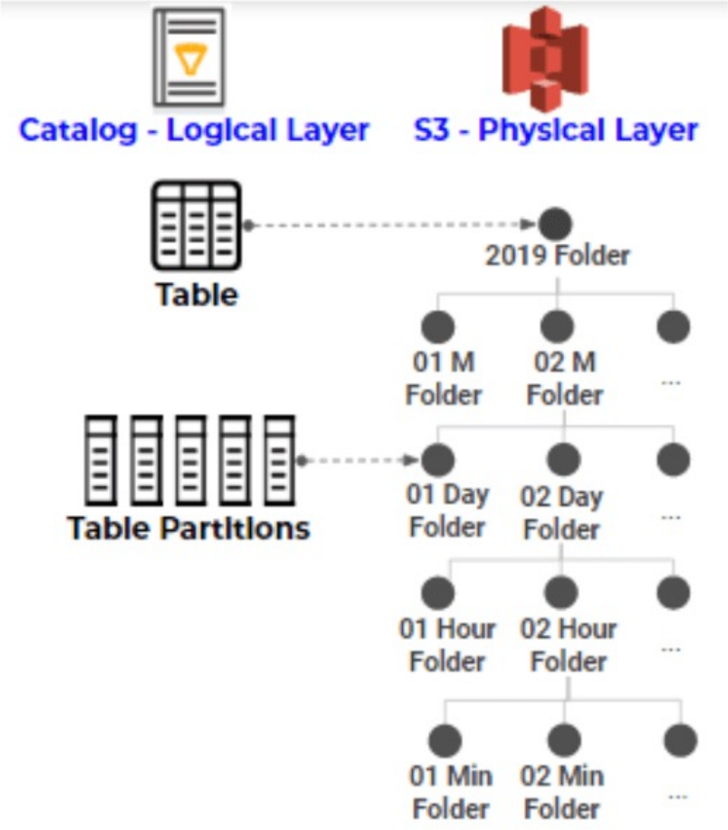
Note the files without Extension is handled due to a property set in the scanner, "Treat Files Without Extension As". if this is set to " Parquet", all the files with no extension will be considered as parquet

Individual file(s) detected and cataloged



Partition Detection Logic

Scan/Profile Logical Layer



- Create Logical Layer on Hive (EMR) – which can be pointed to S3 FileSystem (Parquet, Avro, CSV etc.)
- If there is Complex Data Type like struct, array, break it down to individual columns using qualifiers or create Hive Views
- File Structure needs to be in Hive Style Partition
- Recommended Execution is on Spark Mode
- Best Practice is to leverage Random Percentage Sampling
- SQL Query interface provides better user experience, there is no staging layer involved

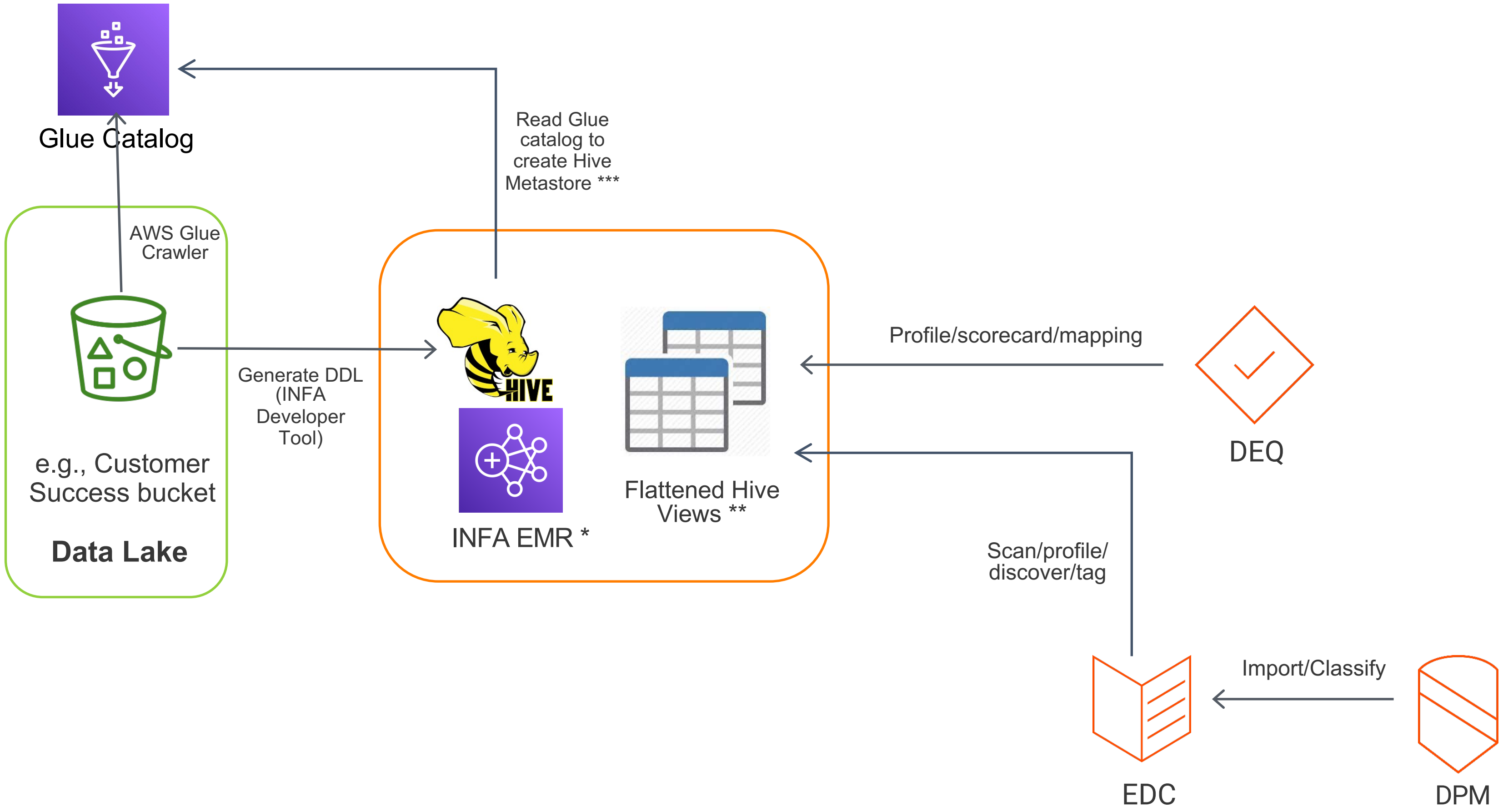
Scan (Options for S3)

Serial number (Ranked)	Approach	Scan Outcomes	Supported Profiling Modes (EDC)	Limitations	DEQ Impact and Profiling Modes
1	Hive on S3 (abstraction layer on S3)	- EDC scan captures metadata including constituents of complex data types **(if flattened)** - Domain discovery and column profiling are performed on all fields including the constituent fields of complex data types **(if flattened)**	- Spark. Random % sampling is supported - Native (not recommended) - Scalable (can enable Auto-Scale on EMR)	- File structure needs to be flattened out first to capture metadata from complex data types. Operation could have overhead and need maintenance - If there is schema evolution, ensure compatible schema being used (Glue crawler looks for upto ~70% of compatible schema, else creates separate tables)	- No impact **(if flattened).** Flattened views can be used as-is in DEQ profiling and scorecard generation process - Customized Data Object (CDO) can be leveraged to flatten the structure using dot qualifier in Custom SQL - Spark Profiling Supported - Scalable (can enable Auto-Scale on EMR)
2	S3 resource (Native)	- Successful scans with partition detection for static S3 schema only - Columns with complex data types are broken down into constituent elements	- Spark. Parquet files only - Supports File Structure with Hive Style Partitioning - Native execution for all other file types	- Schema evolution is not supported, Compatible schemas on the roadmap. - Partition detection and grouping only currently available for parquet files on S3 - Spark pushdown for parquet files only	- Partition detection/grouping not available for S3 files - Item on roadmap but ETA not available - Spark Profiling Supported
3	Athena via jdbc scanner (abstraction layer on S3)	- Performant metadata scan - Performant Domain discovery based on only metadata pattern match	- Native (no sampling)	- Not a natively supported connectivity - Complex data types are not broken down into constituent columns - EDC Data Profiling impacted due to Athena latency. Domain discovery and column profiles impacted - Sampling is not supported for profiling. May cause concern for large datasets	- Native profiling only

Informatica

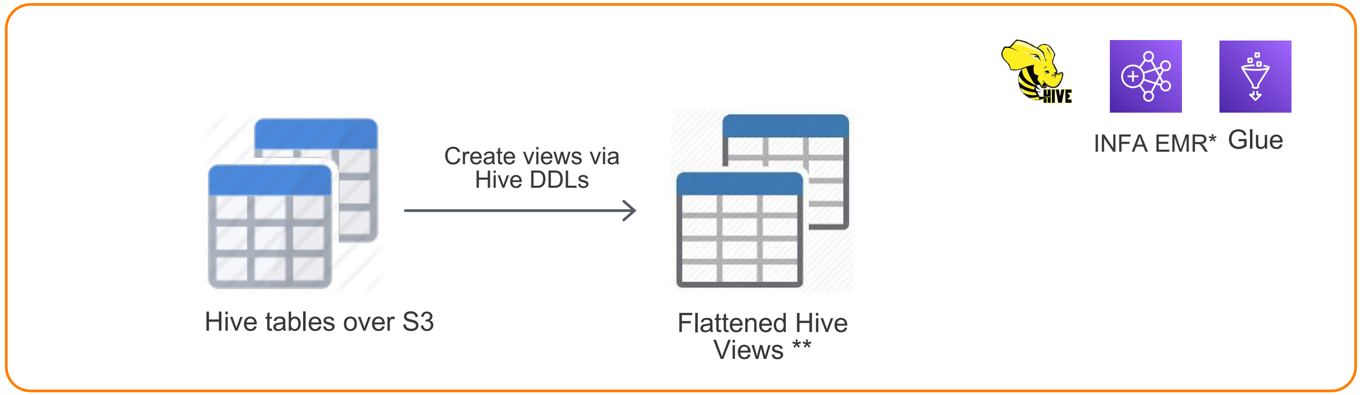
Data Source Prep

Hive on S3 (EMR)


Glue Catalog
Read Glue catalog to create Hive Metastore ***
AWS Glue Crawler
e.g., Customer Success bucket
Data Lake
Generate DDL (INFA Developer Tool)
HIVE
INFA EMR *
Flattened Hive Views **
Profile/scorecard/mapping
DEQ
Scan/profile/discover/tag
Import/Classify
EDC
DPM

** Persistent EMR Cluster used by DEQ and EDC*
*** Flattened Hive views pointing to Hive tables pointing to S3 files. Please see next slide for more info*
**** If EMR is Kerberos enabled, Glue integration has limitations, on the roadmap*

Data Source Prep



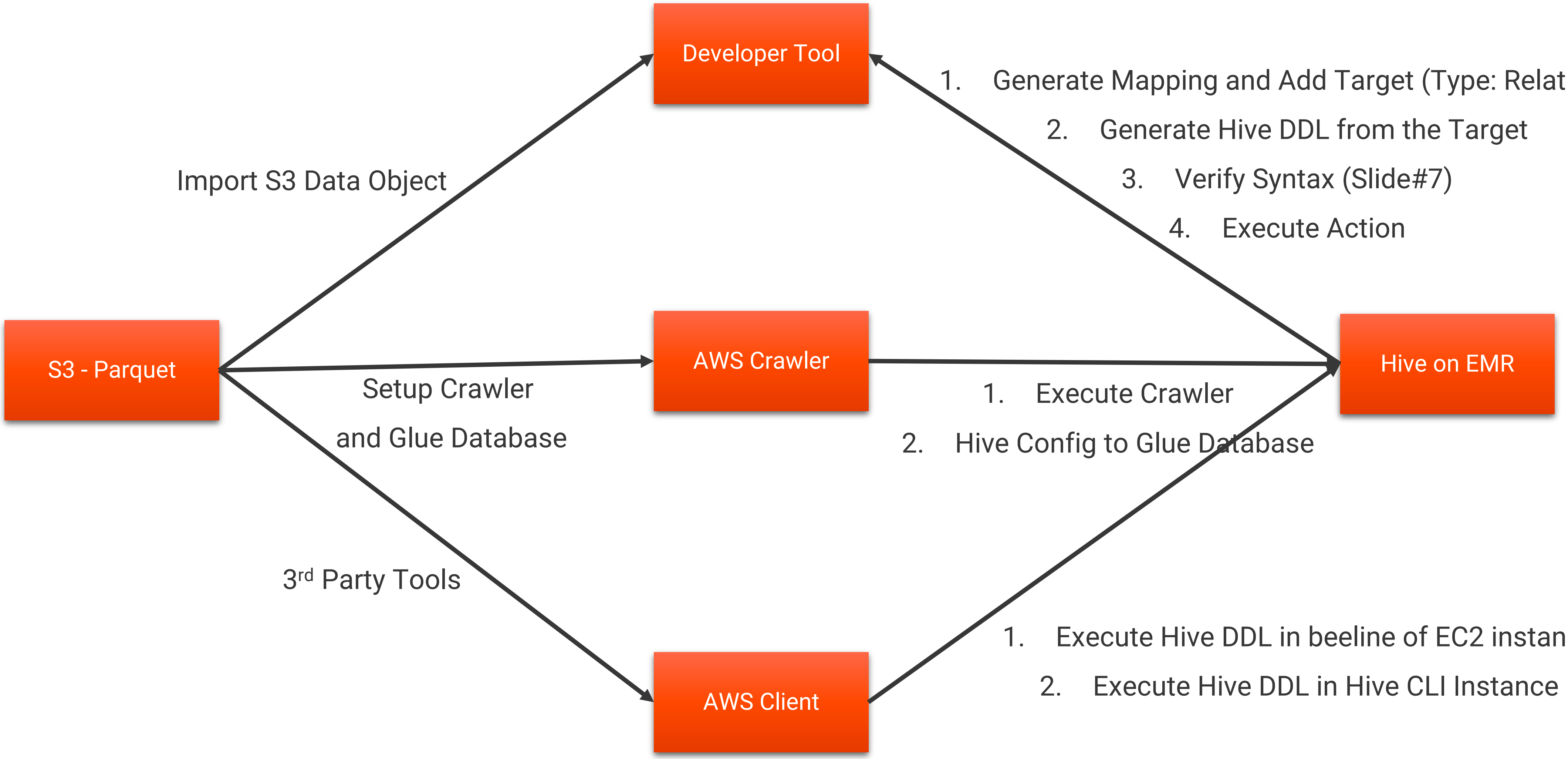
Additional documentation is provided on data source prep tactical steps

Steps	Purpose
Create a database on Hive with supported naming convention	Separate database(s) needs to be created to house flattened views. This database should not have '-'s (hyphens) in the name. Scans will fail if naming conventions are not followed
Create Flattened Hive views over Hive tables	This step is performed in order to *expand fields which uses complex data types*, into its constituent fields. This allows EDC and DEQ to read and process the constituent field values for data quality verification and domain tagging in EDC

Considerations: *please refer to the EMR documentations on considerations when implementing the above approaches*

https://docs.aws.amazon.com/emr/latest/ReleaseGuide/emr-hive-metastore-glue.html

How to create Hive Tables from S3



Import S3 Data Object

Setup Crawler and Glue Database

3rd Party Tools

1. Generate Mapping and Add Target (Type: Relational)
2. Generate Hive DDL from the Target
3. Verify Syntax (Slide#7)
4. Execute Action

1. Execute Crawler
2. Hive Config to Glue Database

1. Execute Hive DDL in beeline of EC2 instance
2. Execute Hive DDL in Hive CLI Instance

Hive on S3

Generate DDL using Developer Tool

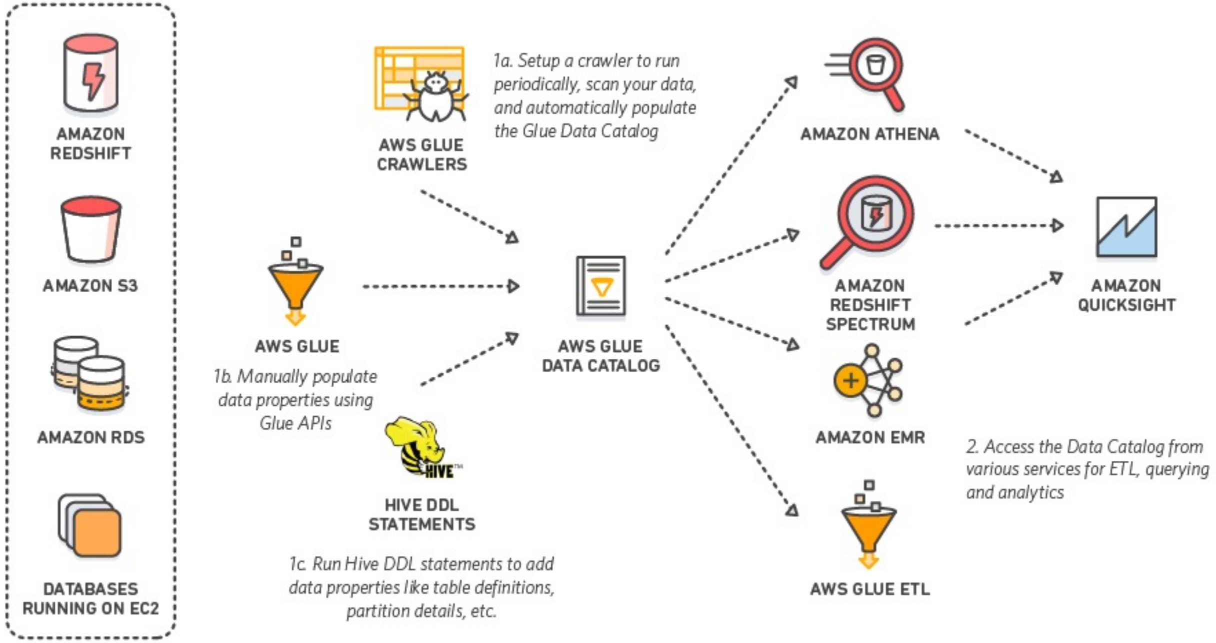
Syntax

```
CREATE EXTERNAL TABLE [IF NOT EXISTS] [db_name.] table_name
[(col_name data_type [COMMENT col_comment], ...)]
[COMMENT table_comment]
[ROW FORMAT row_format]
[FIELDS TERMINATED BY char]
[PARTITIONED BY column datatype]
[STORED AS file_format]
[LOCATION S3_Path];
```

Example

```
CREATE EXTERNAL TABLE sensor
(
room string,
energy double,
temp double,
occupancy int,
awhen timestamp
)
PARTITIONED BY (year string, month string, day string)
STORED AS PARQUET
LOCATION 's3://s3.us-west-2.amazonaws.com/a2g-hive-
test/tempsensores/data/';
```

AWS Glue Crawler


AMAZON REDSHIFT
AMAZON S3
AMAZON RDS
DATABASES RUNNING ON EC2
AWS GLUE CRAWLERS
1a. Setup a crawler to run periodically, scan your data, and automatically populate the Glue Data Catalog
AWS GLUE
1b. Manually populate data properties using Glue APIs
HIVE
HIVE DDL STATEMENTS
1c. Run Hive DDL statements to add data properties like table definitions, partition details, etc.
AWS GLUE DATA CATALOG
AMAZON ATHENA
AMAZON REDSHIFT SPECTRUM
AMAZON EMR
AWS GLUE ETL
AMAZON QUICKSIGHT
2. Access the Data Catalog from various services for ETL, querying and analytics

Maintenance

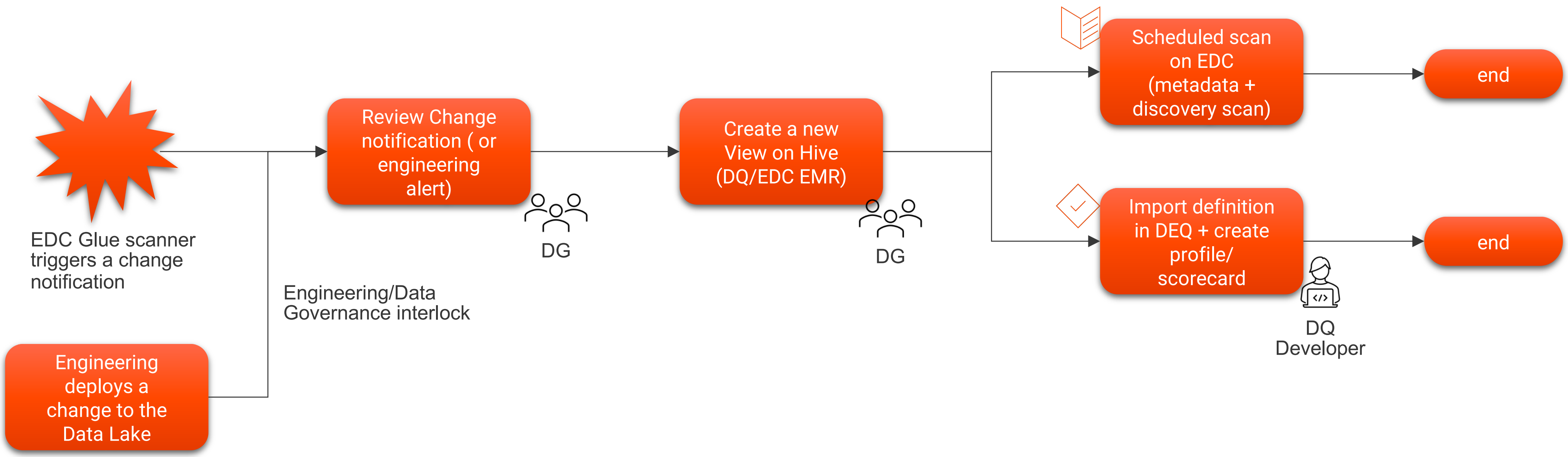
Maintenance of Source Metadata

- *Additions*
  - New views on Hive need to be created as new buckets are added into the scope for scanning
  - Keep track of changes on existing table structures using the “change notifications” on Glue scanner

- *Backups*
  - *No special backups of HDFS or Hive are needed for the views created on EMR. The view definitions are stored in the Glue metastore, is completely managed by AWS. No information is locally managed on HDFS*

Maintenance Process Flow

Addition of a new table (sample)



EDC Glue scanner triggers a change notification

Engineering deploys a change to the Data Lake

Engineering/Data Governance interlock

Review Change notification (or engineering alert)

DG

Create a new View on Hive (DQ/EDC EMR)

DG

Scheduled scan on EDC (metadata + discovery scan)

end

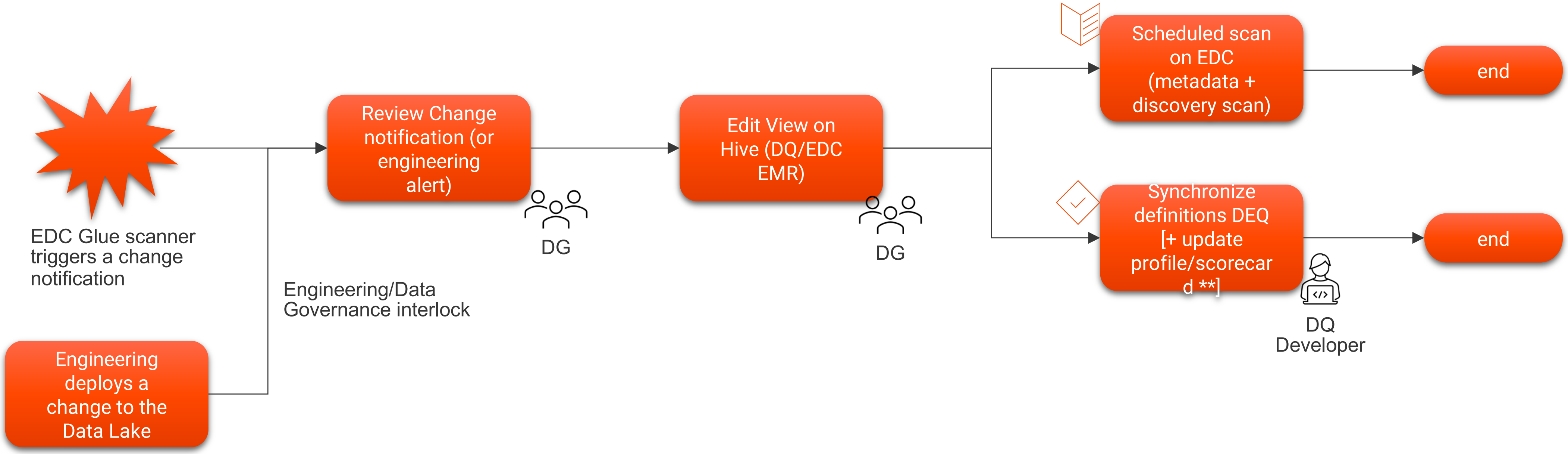
Import definition in DEQ + create profile/ scorecard

DQ Developer

end

Maintenance Process Flow

Addition of a new column in an existing table (sample)



EDC Glue scanner triggers a change notification

Engineering deploys a change to the Data Lake

Engineering/Data Governance interlock

Review Change notification (or engineering alert)

DG

Edit View on Hive (DQ/EDC EMR)

DG

Scheduled scan on EDC (metadata + discovery scan)

end

Synchronize definitions DEQ [+ update profile/scorecard **]

DQ Developer

end

***update profile/scorecard if the new field needs to be validated through a DQ check*

Scanner Configuration

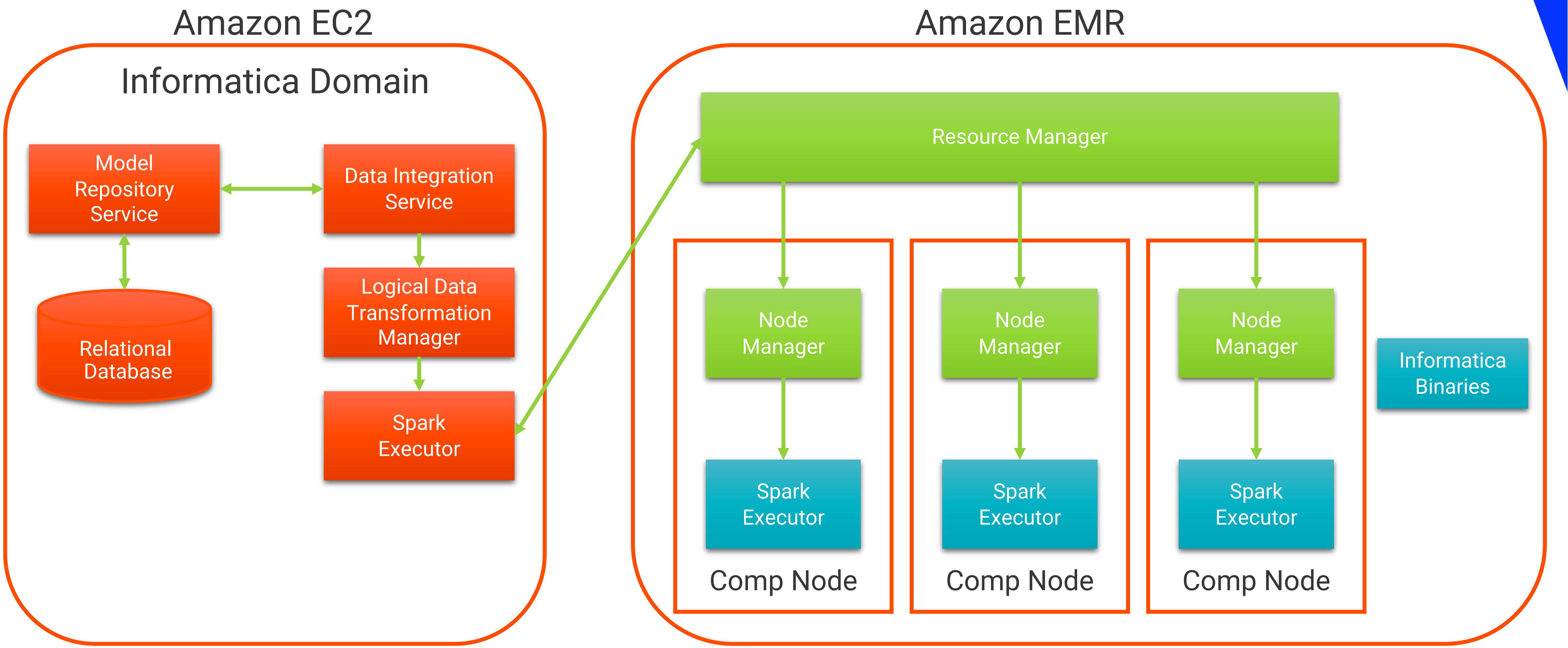
EDC Scan Config (EMR)

Source Type: Hive on S3

Scan Type	Resource/Connection Type	Connection Management UI	Pros/Cons	**Pre-requisites** (* sans server details and credentials)
Metadata	jdbc	Catalog Administrator	If you have Hive Views created, its Best Practice to use JDBC Resource Type, else Hive Resource Type need to be used. Note: All the views are treated as tables in Hive	Install hive jdbc drives on EDC servers
Profiling/Domain Discovery	Hive Native Connection	EDC Admin Console	Pros: Native Hive connection enables spark pushdown profiling. As a best practice choose "Random Percentage" as sampling option for optimum performance and best results.	None

Infrastructure

Spark Processing



- 🟠 Informatica Domain Services
- 🔵 Informatica Hadoop Binaries
- 🟢 Hadoop Services

EMR Cluster Recommendation

Refer to Informatica Product Availability Matrix for supported versions on EMR

Services Needed

- Glue
- Hive
- Hadoop

Configurations

- Autoscaling – ensure autoscaling is enabled to handle varying degrees of execution loads for EDC and DEQ

Limitations

- Kerberos with Glue on EMR is not certified, on roadmap
- DEQ/EDC requires a persistent EMR cluster (for profiling). Cluster may be temporarily stopped when not in use but should not be terminated.

***Considerations:** please refer to the EMR documentations on considerations when implementing the above approaches*

https://docs.aws.amazon.com/emr/latest/ReleaseGuide/emr-hive-metastore-glue.html

Frequently Asked Questions

1. DEQ and EDC using same EMR cluster, what will be the impact on performance with both services consuming? Can we reserve specific nodes/resources for each product?
- Sized EMR appropriately provides Elasticity (Scale Out)
- Scheduling Frequency
- YARN Job Queues can be used to Optimize the cost, workload management

2. Can the EMR can be used as Ephermal cluster?
- Its possible with Mappings
- With Profiling its not certified (Needs to be scripted in AWS)
  Profiling through infacmd, part of the workflow, Pre Script
- External Hive Metastore incase EMR Cluster is killed

3. As customer is looking for Self Service Model – Is there a way to quickly generate Hive Tables based on S3 files?

Here are the options:

- Hive DDL generation tools from S3 Parquet
- Scan through EDC, get the File Structure and generate DDL using REST API
- AWS Crawler, Crawler has constraints like Table Names cannot be renamed etc

https://docs.aws.amazon.com/emr/latest/ReleaseGuide/emr-hive-metastore-glue.html

Frequently Asked Questions

4. How does Hive work with S3 from Security perspective?
As per AWS Support: Apache Hive runs on Amazon EMR clusters and interacts with data stored in Amazon S3. Hive runs on top of Hadoop, with Apache Tez or MapReduce for processing and HDFS or Amazon S3 for storage. For example AWS S3 will be used as the file storage for Hive tables. A session token can be used to provide temporary credentials that provide the same permissions that you have with use long-term security credentials such as IAM user credentials. Hive should by default use instance profile and it will take care of IAM credentials and tokens configured. These credentials are then used to make a call to Amazon S3 when needed. Basically an AWS Account or an IAM user can request temporary security credentials and use them to send authenticated requests to Amazon S3

References

1. Specifying AWS Glue as Hive Metastore: Click here
2. Integration with AWS Glue Data Catalog: Click here
3. Setting up Hadoop Cluster Configuration in DEQ with Amazon EMR: Click Here
4. EMR Sizing Guidelines: Click Here
5. Scanning DataLake Filesystems: Click Here

Q&A

Thank You