



CAPGEMINI RESPONSE TO TELIA – CDL TO ARCUS

15th July 2024

Telia has defined an architecture for new data lake platform (i.e Arcus) to replace CDL2 with migration of data and corresponding connection from source by providing similar experience to consumer applications (like Olik , export)

Replace of CDL2 is to address following objectives

- CAPEX & OPEX reduction
- License cost optimization by Sept 2025
- GDPR compliance

Arcus platform components are integrated, and platform is ready for adding sources with Data pipeline

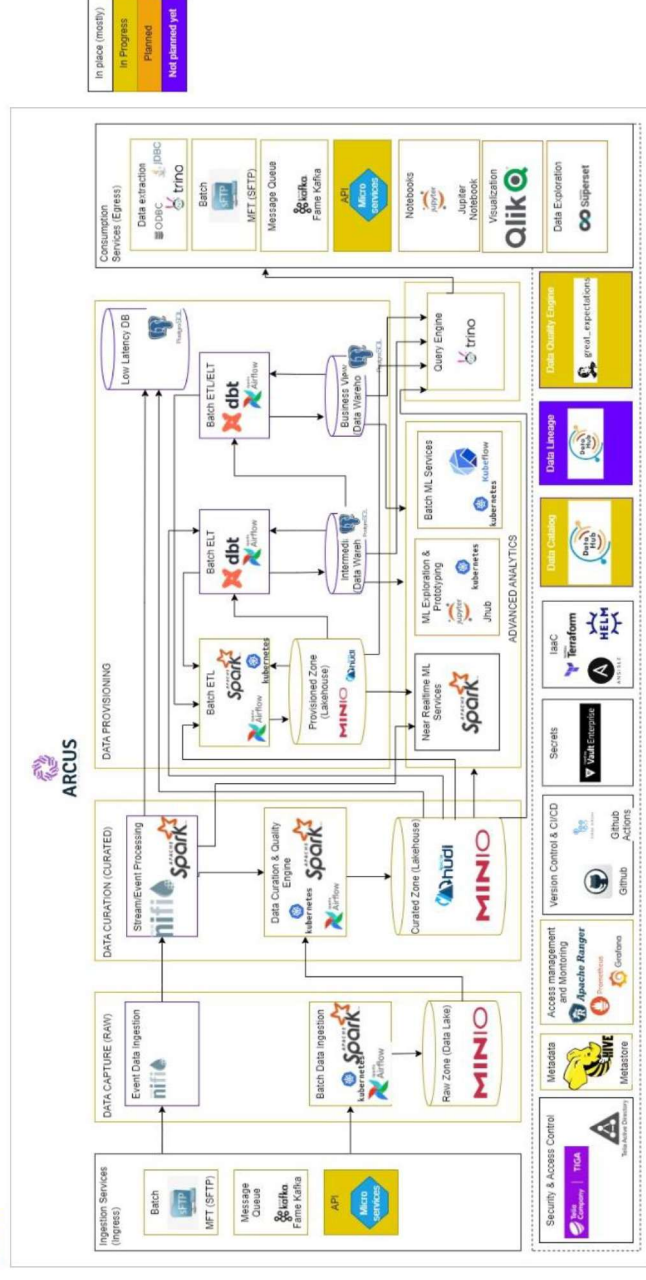
- Country – Sweden only (B2C,B2B,Martec) + Group Analytics
- There are certain sources which will get directly push to S3 directly (certain sources are not allowed to access to offshore)
- Telia has already done analysis for data sources which can be used for planning the migration

Volumetric for the scope combined Sweden and Group Analytics as follows

Analytics as follows

- ~150 Sources
- ~833 (+5%) objects in scope (1389 total in source system)
- ~583 objects in curated layer
- ~163 object in provisioning layer (reduced after inputs from Telia)

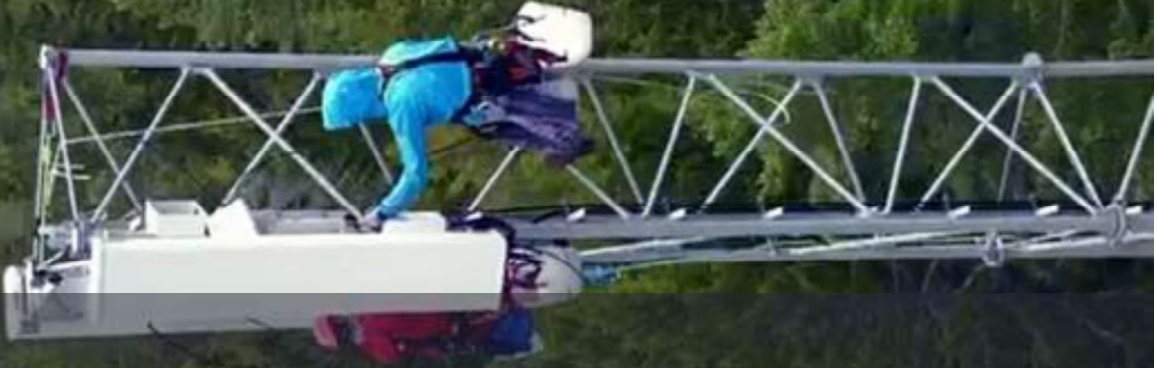
*Based on Discover findings we will revisit estimation and Wave plan



Architecture provided by Telia subject to revisions in Technical components in future



OUR APPROACH





SCOPE COVERAGE PER WAVE FOR SWEDEN B2C,B2B & MARTEC

Data Ingestion & Processing

- Define workflows for data ingestion, transformation, validation and loading jobs, Business rules, Routine and schedules
- Define data pipeline and ETL jobs for data ingestion , curation and Provisioning for Sweden
 - B2C (30%) , B2B (50%) , Martec (20%)
 - For B2C, data format conversion to Parquet Data format as data moving to S3 directly
 - Volume of data movement Raw → Curated (10% reduction) , Raw → provisioning (40%)
- Data migration for B2C , B2B and Martec with no change in existing data model
- Implement Telia provided data model for the new data repository
- History Data (~12 TB) load to new data repository

Data Quality

- Define data quality technical framework and KPIs.
- Implement Data Quality checks on GDPR and Personal Information(PI) related attributes
- Data masking and encryption
- Use of possible tool Great expectation and AWS Deequ
- Data profiling and score card

Data Governance Strategy

- Integration with existing Data Hub
- Follow Telia defined Data Governance strategy and principles

Reporting

- Query engine Trino need to be use for all data consumption
- Reports Rationalization to finalize the number of reports and dashboards to change connection settings towards Trino s part of Qlik Apps and Export files
- Validation of KPIs and reports

General:

- Knowledge acquisition of current platform from Telia and Identify areas of improvement
- Documentation of Parallel data layer(Data Schema, data flow diagrams, blueprints and data source metrics)
- SIT and UAT Support for the in-scope components
- Virtual KT and trainings workshops
- Knowledge transfer to Telia Operations team for maintaining and upgrading the data platform solution



SCOPE COVERAGE PER WAVE FOR GROUP ANALYTICS

Data Ingestion & Processing

- Define workflows for data ingestion, transformation, validation and loading jobs, Business rules, Routine and schedules
- Define data pipeline and ETL jobs for data ingestion and curation for Group Analytics
 - Volume of data movement Raw → Curated (10% reduction) , Raw → provisioning (40%)
- Data migration with no change in existing data model
- Implement Telia provided data model for the new data repository
- History Data (~5 TB) load to new data repository

Data Quality

- Define data quality technical framework and KPIs.
- Implement Data Quality checks on GDPR and Personal information(PI) related attributes
- Data masking and encryption
- Use of possible tool Great expectation and AWS Deequ
- Data profiling and score card

Data Governance Strategy

- Integration with existing Data Hub
- Follow Telia defined Data Governance strategy and principles

Reporting

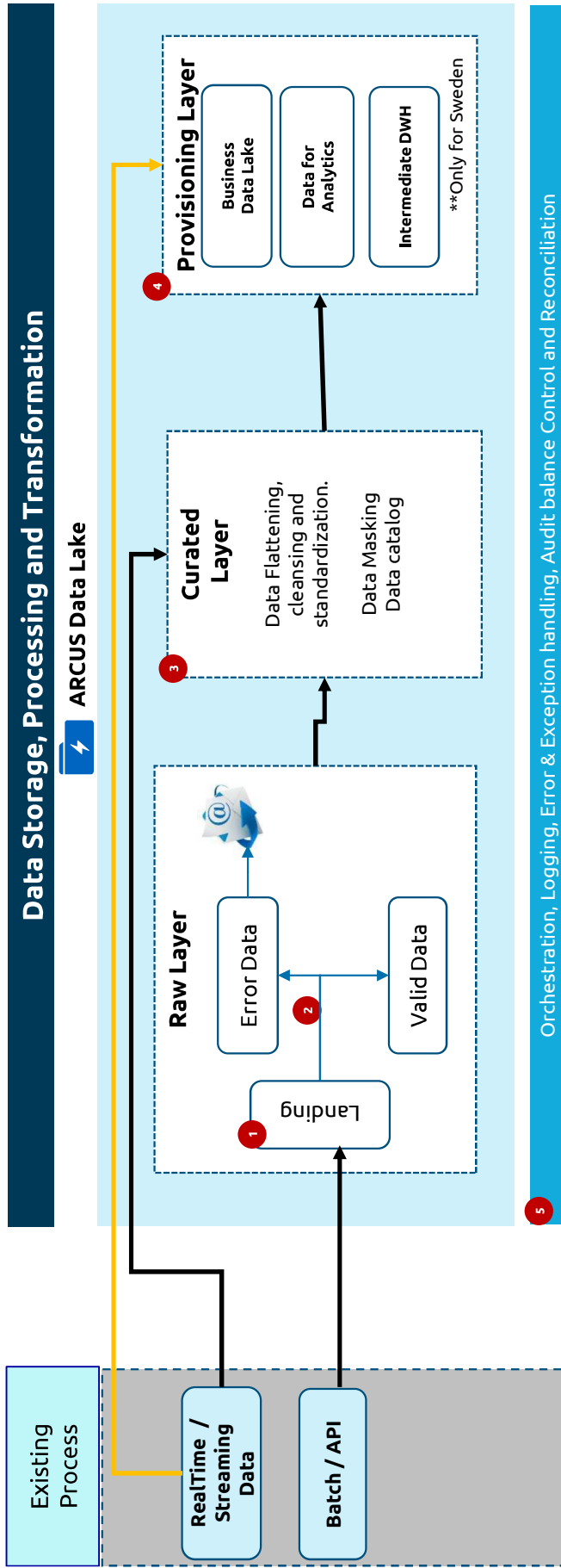
- Not in scope for Group Analytics

General:

- Knowledge acquisition of current platform from Telia and Identify areas of improvement
- Documentation of Parallel data layer(Data Schema, data flow diagrams, blueprints and data source metrics)
- SIT and UAT Support for the in-scope components
- Virtual KT and trainings workshops
- Knowledge transfer to Telia Operations team for maintaining and upgrading the data platform solution



OUR UNDERSTANDING OF BUSINESS DATA FLOW AND ACTIVITIES - DATA INGESTION AND PROCESSING



1

Data will be extracted from existing data source. Data will be loaded as-is into landing layer of data lake using mass ingestion. This will include streaming and batch data load

2

Data from landing area will be processed into Error and valid folders of raw layer while performing technical file validation. notification will be sent for error data.

3

Data from valid logical folder of raw layer will be processed into curated layer. Data will be flattened, cleansed, standardized. All the CDE attributes will be mask
Data glossary – dictionary will be prepared

4

Business rules will be applied on data from curated layer and processed into enriched layer for Data Warehouse and analytical use cases

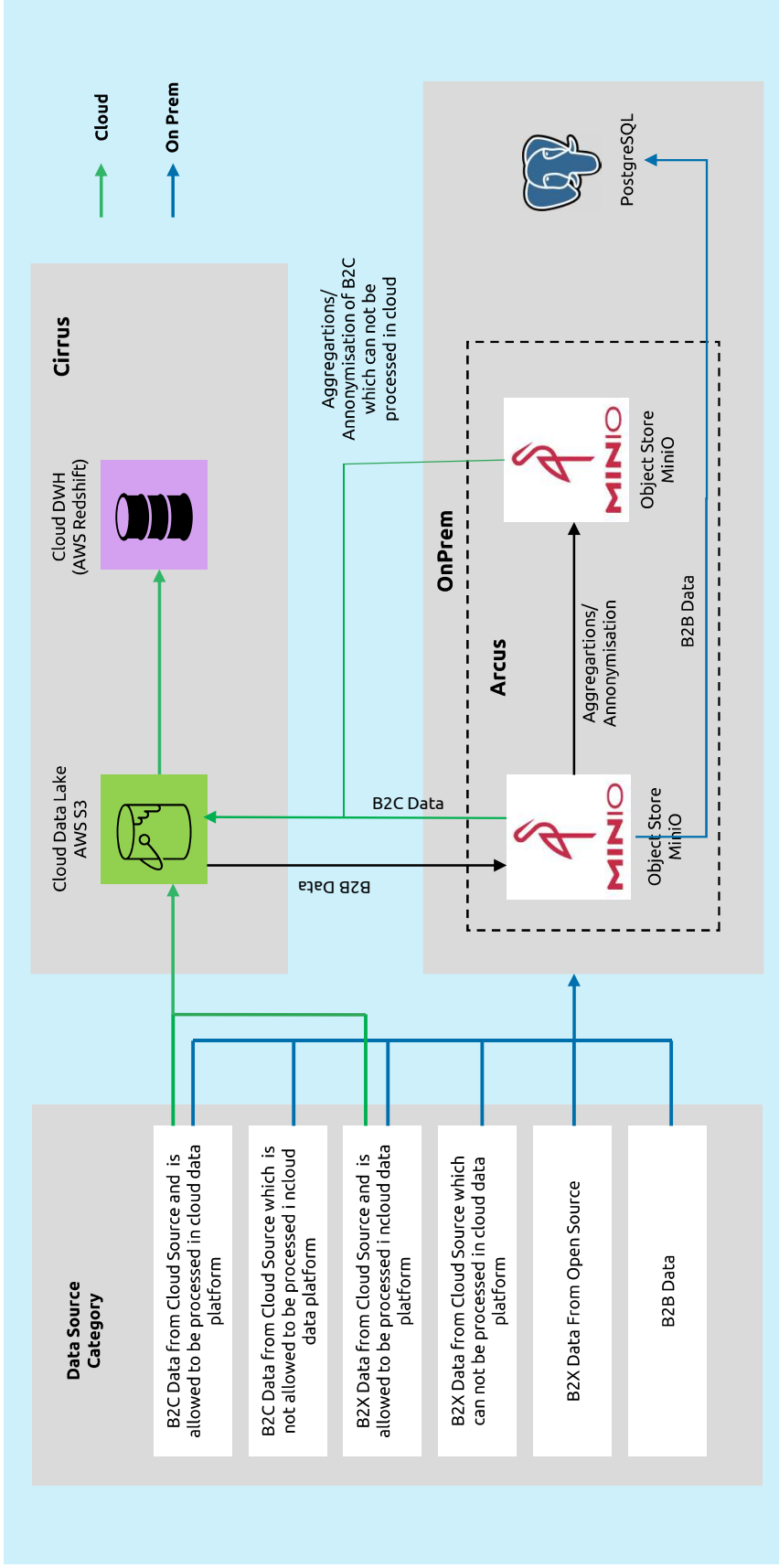
5

Orchestration, logging, error & exception handling, audit balance control and reconciliation will be implemented for ETL jobs



HIGH-LEVEL TARGET DATA FLOW FOR VARIOUS DATA CATEGORIES WITHIN SWEDEN

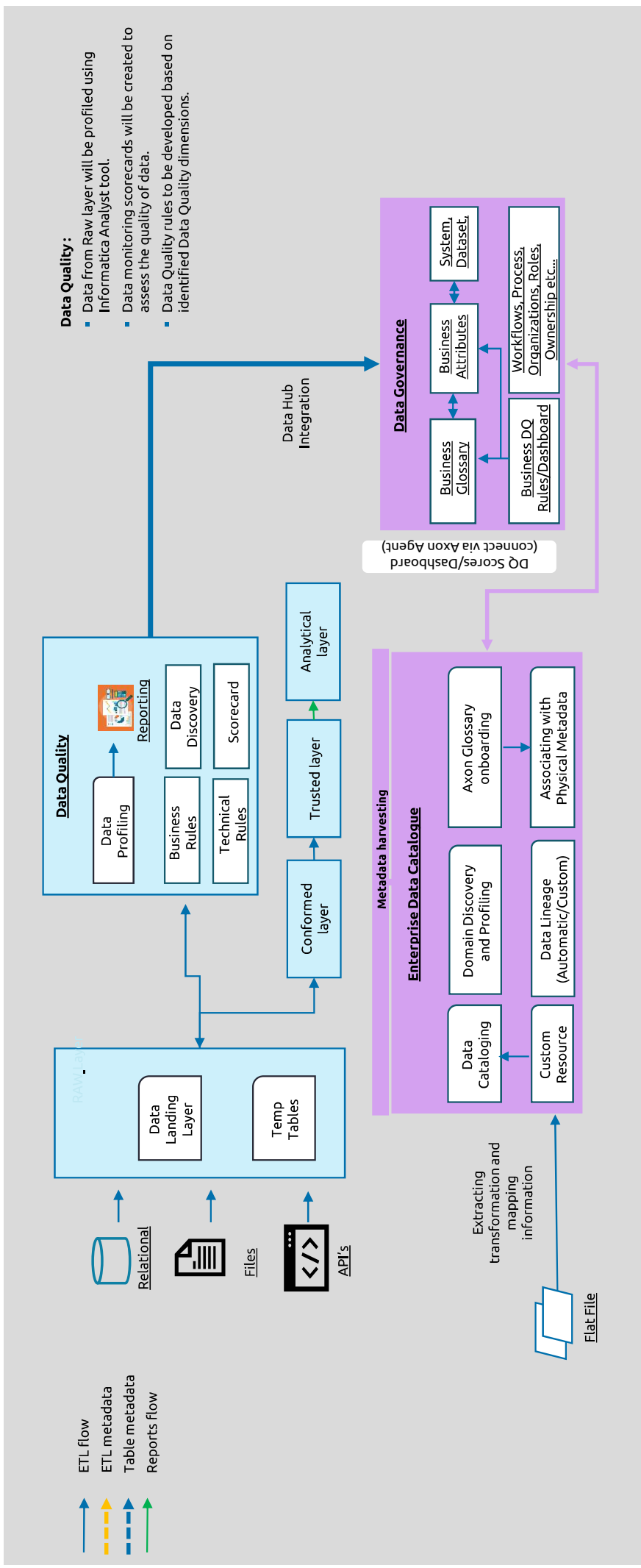
** Reference Telia defined strategy*



As the target platform is already built , our scope is limited to data ingestion into MINIO whereas B2C data will move to Redshift DWH.



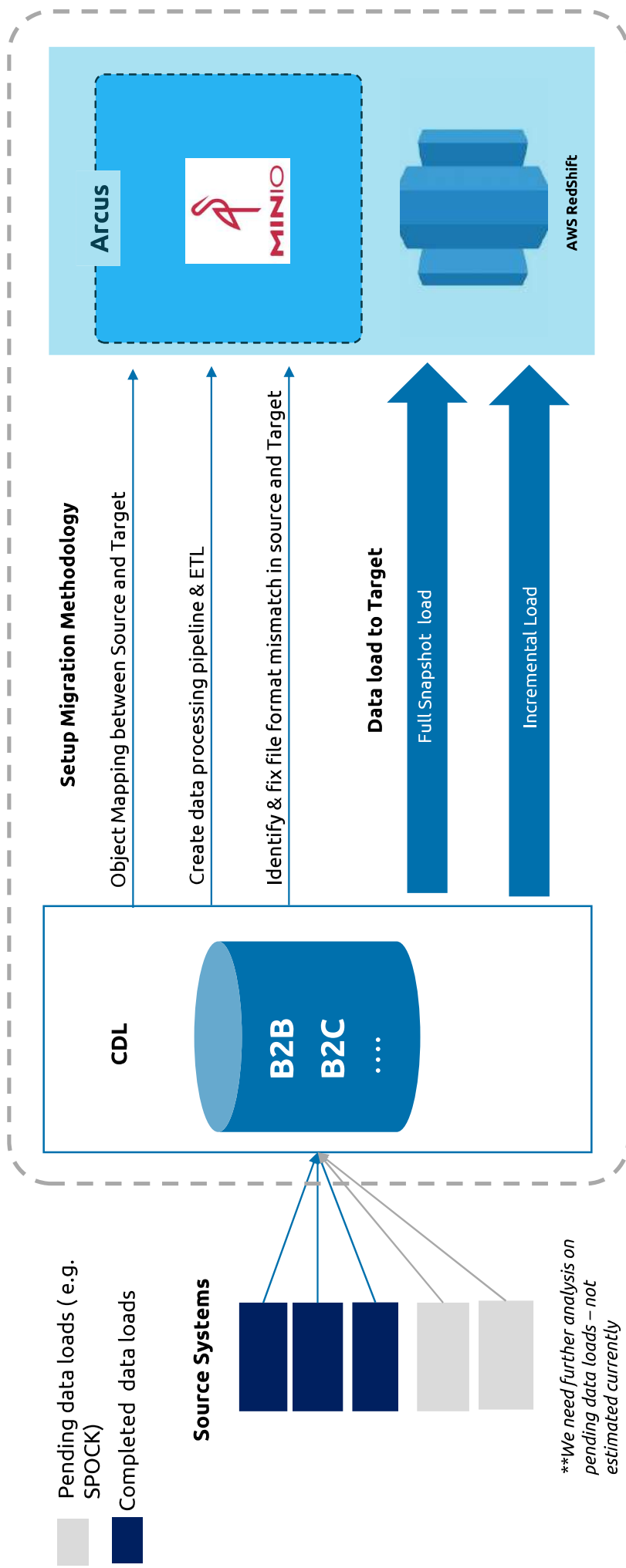
DATA QUALITY AND DATA GOVERNANCE IMPLEMENTATION



As shown above the best practices from Capgemini will be followed , with understanding of all Data Governance configurations will be done inside DataHub and our scope is limited to integration towards Datahub. As part of DQ will perform the listed activities



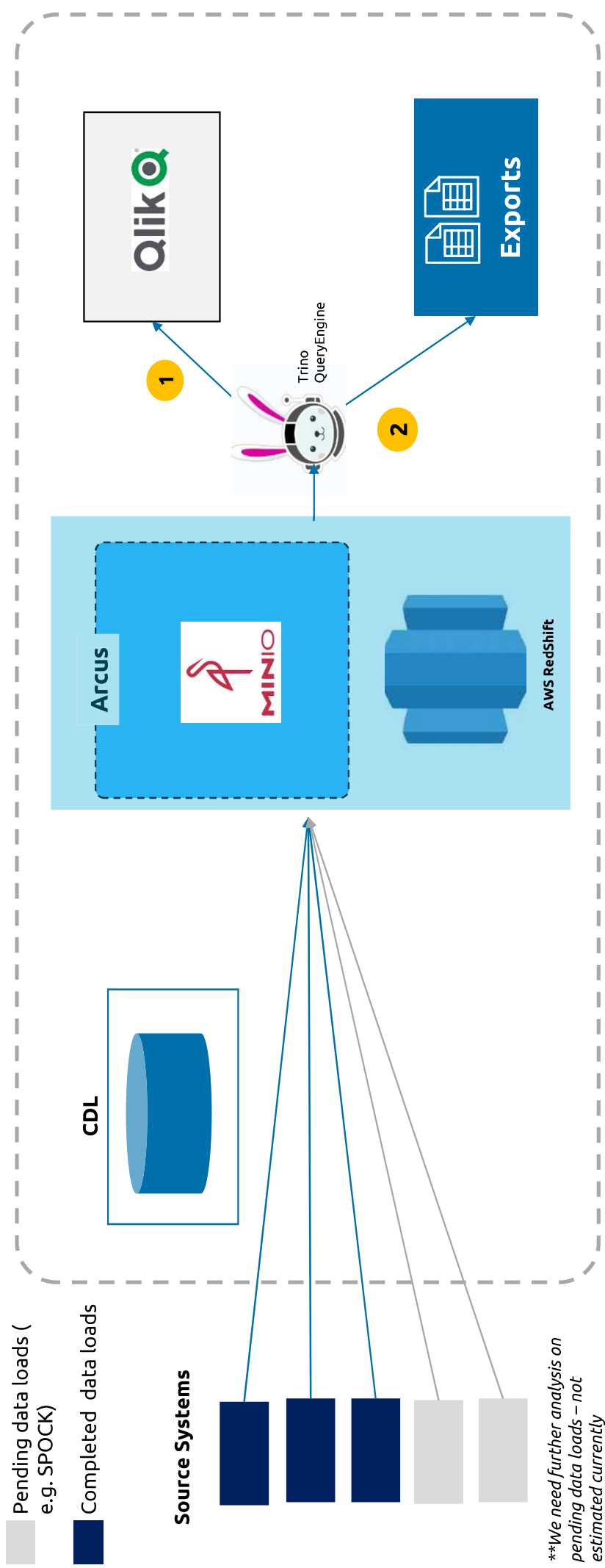
ACTIVITIES AS PART OF MIGRATION FROM CDL2 TO ARCUS/REDSHIFT



As part of migration scope we have considered data model and structure are already defined (in alignment to CDL2) the activities involved are as shown above



ACTIVITIES AS PART OF CONSUMPTION (REPORTS, EXTRACTS)



- 1 • Connect changes towards New Platform
- 2 • Connect changes towards New Platform
- Report KPI Validation
- File Validation

ASSUMPTIONS AND DEPENDENCIES



Updated

DEPENDENCIES

- Architecture knowledge sharing and access to documentation
- Availability of Telia Roles for Platform support , UAT , Data Source analysis and validation
- CG Access to Platform and components with Dev, Test and Production Environment
- Timeline Signoff for the Wave specific milestones
- Squad/business will be responsible for highlighting changes needed in Access layer
- Availability of Dev, Test and Prod environments to be secure by Telia as per Wave time plan

ASSUMPTIONS

- Compliances at Data layer to apply (DQ, Right file format for each table – open data format like Iceberg/Hudi)
- For Group Analytics final stage of data will be till Curated layer
- We assume efficiency gain of approx. 25 – 30 % on subsequent Waves
- Data model and Schema for source and target systems are same and will be provided by Telia
- The scope is limited to data validation and repointing of reports. Any bug fixing (other than related to repointing to Trino) or enhancements in existing reports are out of scope.
- Based on Discover findings we will revisit Wave plan
- Capgemini will be responsible for lift and shift migration and any changes in design/architecture will be under Telia's accountability

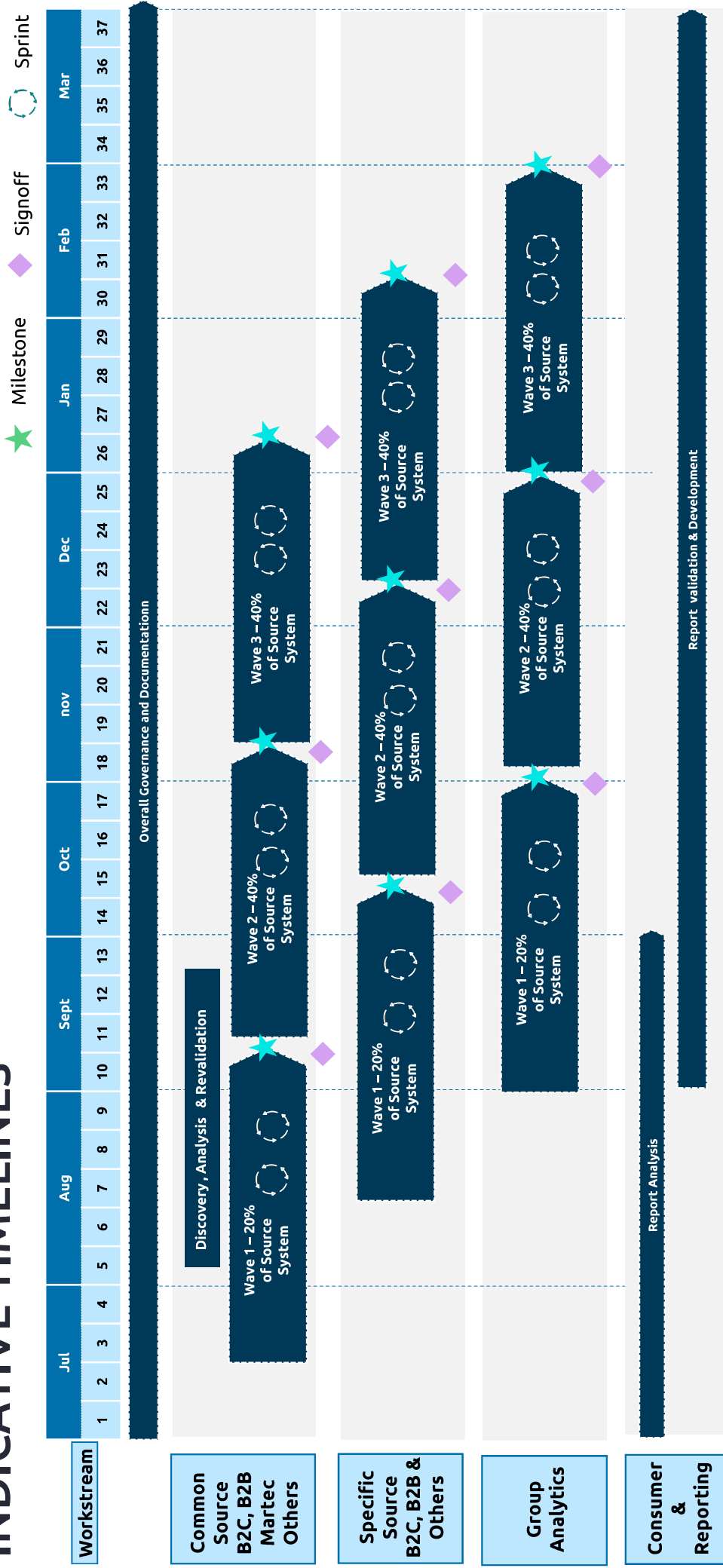


TIMELINES & TEAM





INDICATIVE TIMELINES



***Early start of Waves is possible with understanding that Telia has already done analysis for data sources which can be used for initial waves and for remaining we will do analysis in parallel with validations on our understanding of scope at this stage**



PROPOSED TEAM STRUCTURE

