MX2020 Big Data

Data Leke Reference Architecture

Alerts and Notifications

March 19, 2020

Change control

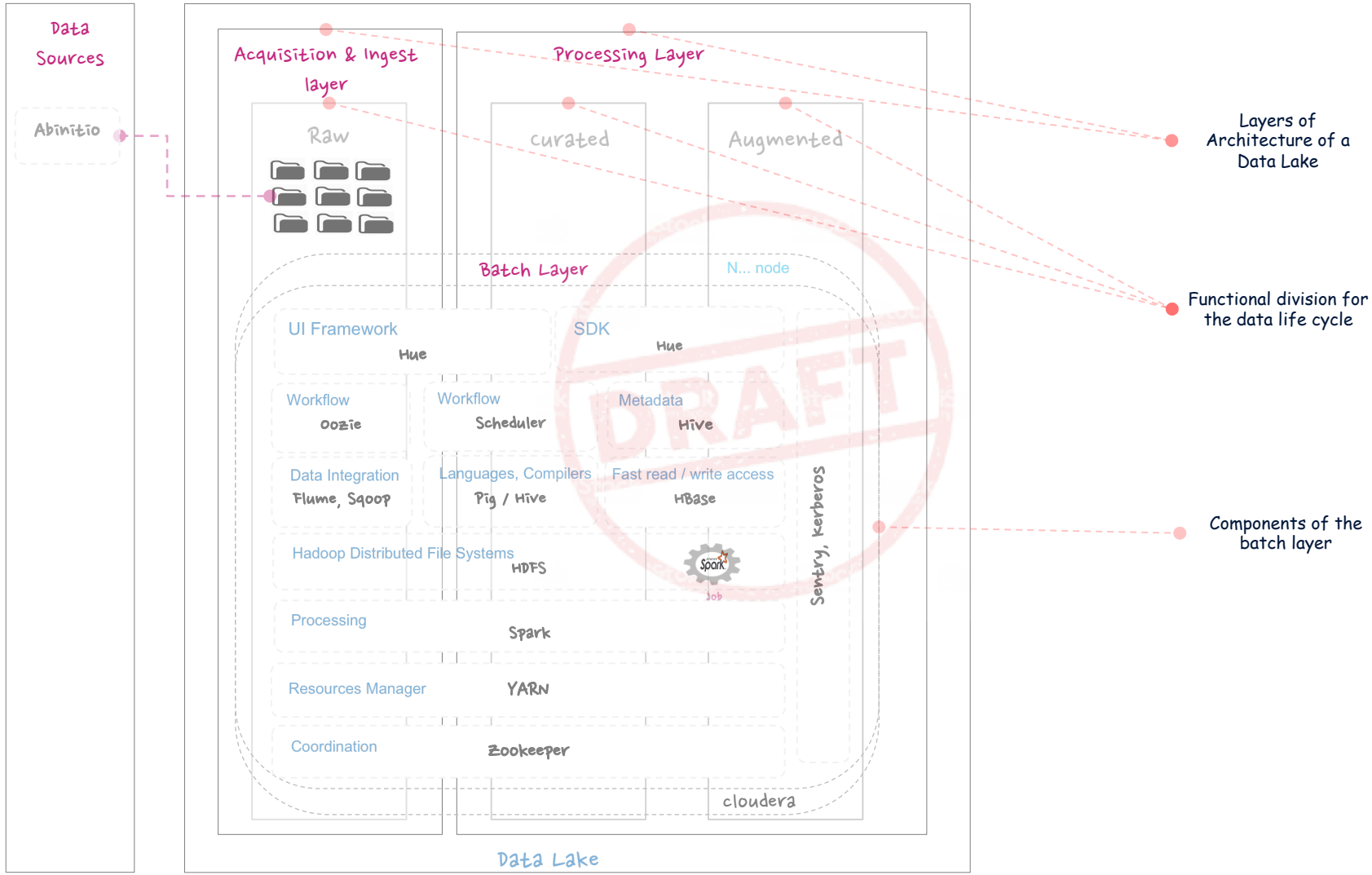
Versión	Fecha	Descripción del Cambio	Autor/Departamento
0.1	10/05/2019	Creation of the document	[Big Data Architecture]

Table of Contents

Features and terminology

Features and terminology (I)

Data Lake - Alerts and Notifications



Needs and Solution Proposal (I)

Descripción de la solución técnica

Objective

- Define core components, pieces of code for the Data Lake Notification and Alerts project architecture

Scope

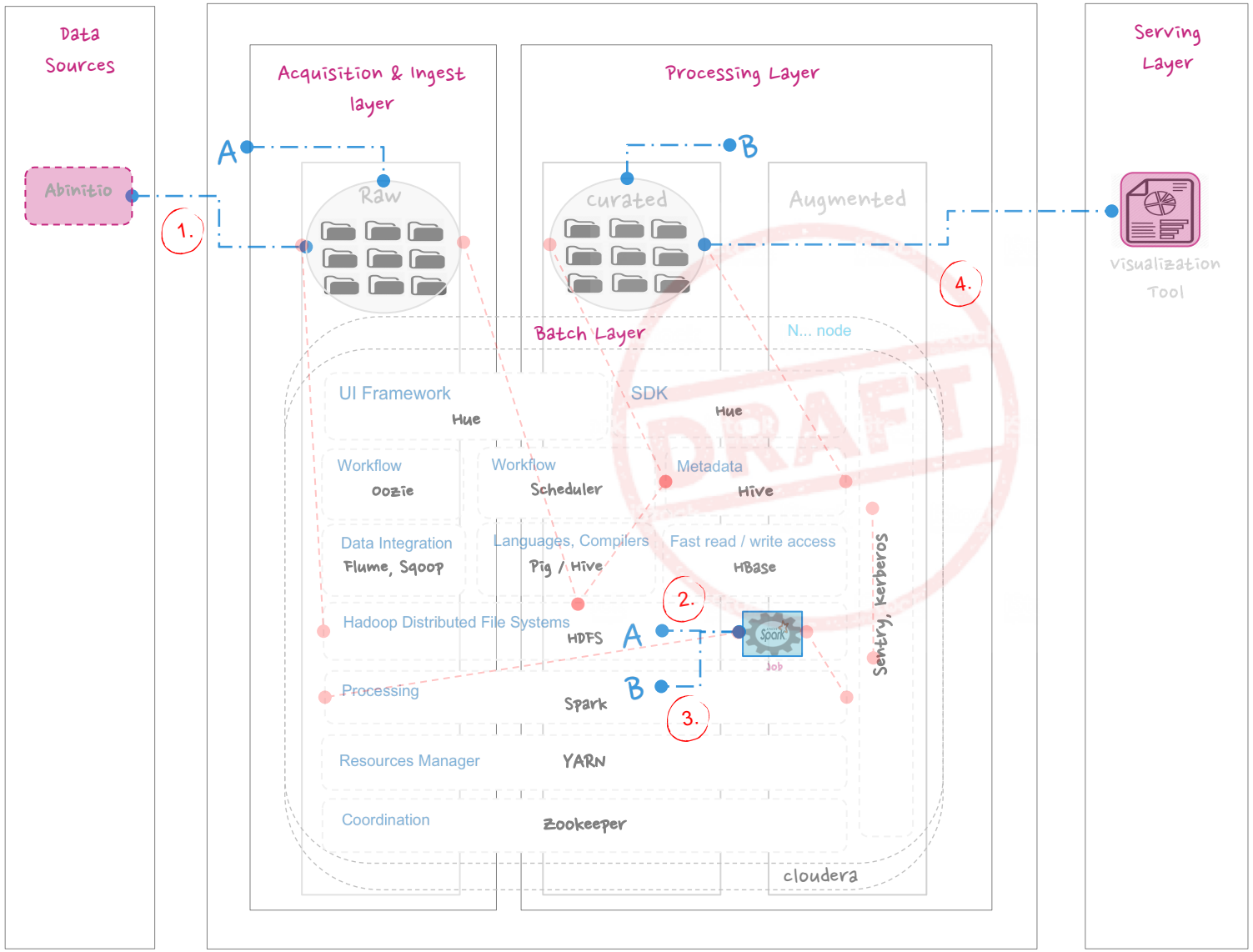
- Represent graphics of the components, internal structure and integrations, incorporating the use of good design and development practices in the different phases of the software engineering process.

Description

- Due to the reengineering of the Architecture of the Alerts and Notifications area, the following functionalities for the Data Lake are identified:
    - Validate the structure of the Event
    - Move Row Zone data to Cured
    - Process the data according to certain Alerts and Notifications business rules, make a view available for the visualization tool
    - Completeness and data quality

Needs and Solution Proposal (II)

Data Lake - Alerts and Notifications – Global Architecture

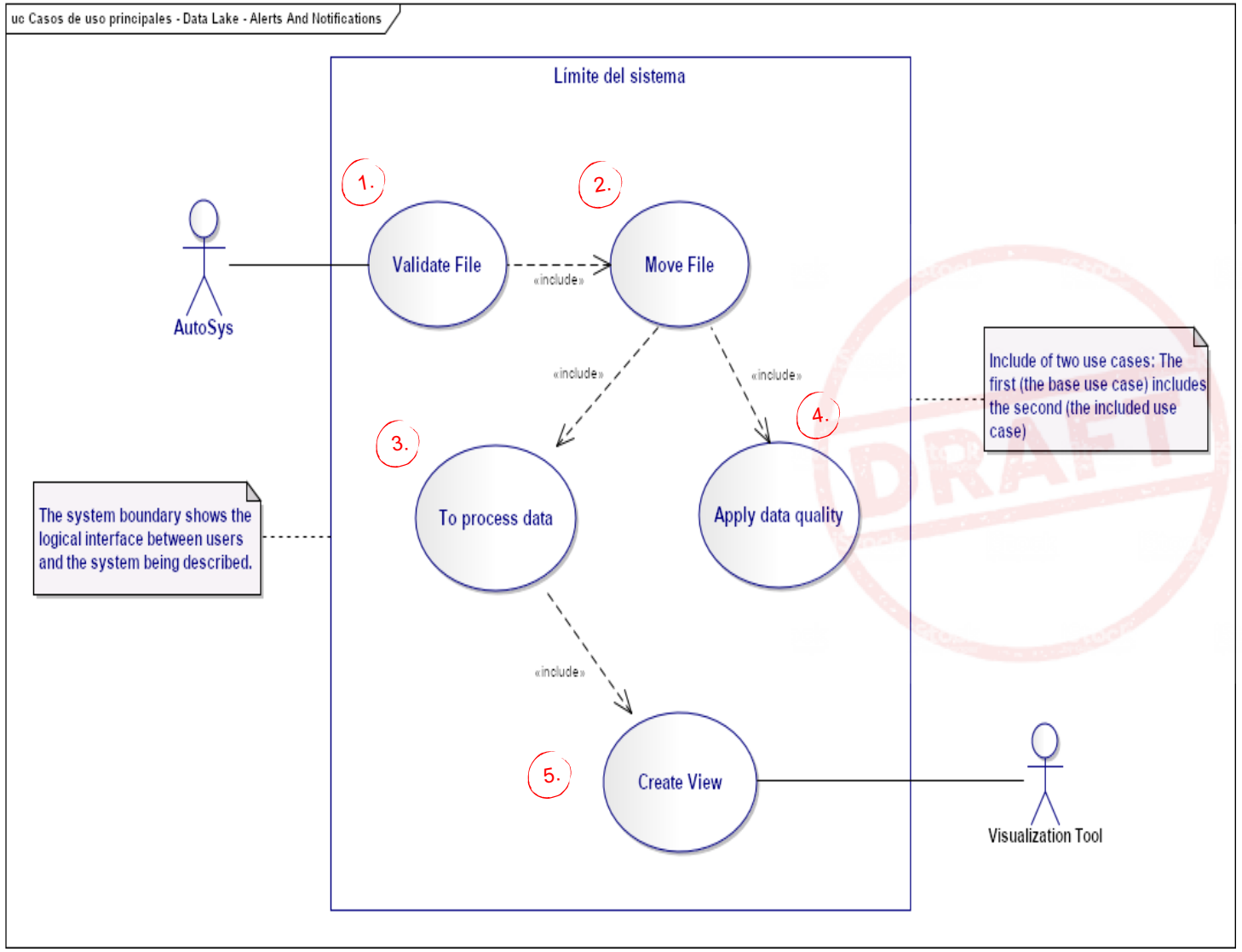


New component
Component to modify
Component not contemplated
General flow
Integration with Hadoop components

1. Abinitio ETL sends the Json file with the events of the day to the Zona Row
2. El Job de Data Lake valida la existencia del archivo y la estructura del Json
3. The Data Lake Job validates the existence of the file and the structure of the Json
4. The Data Lake job moves Row data to cured by creating the table in Hive, processes the information based on the Alerts and Notifications business rules
5. The Data Lake Job creates a view in Hive with the data it processes to make it available to the tool in visualization

Needs and Solution Proposal (II)

Event Hub Architecture – Use Case View

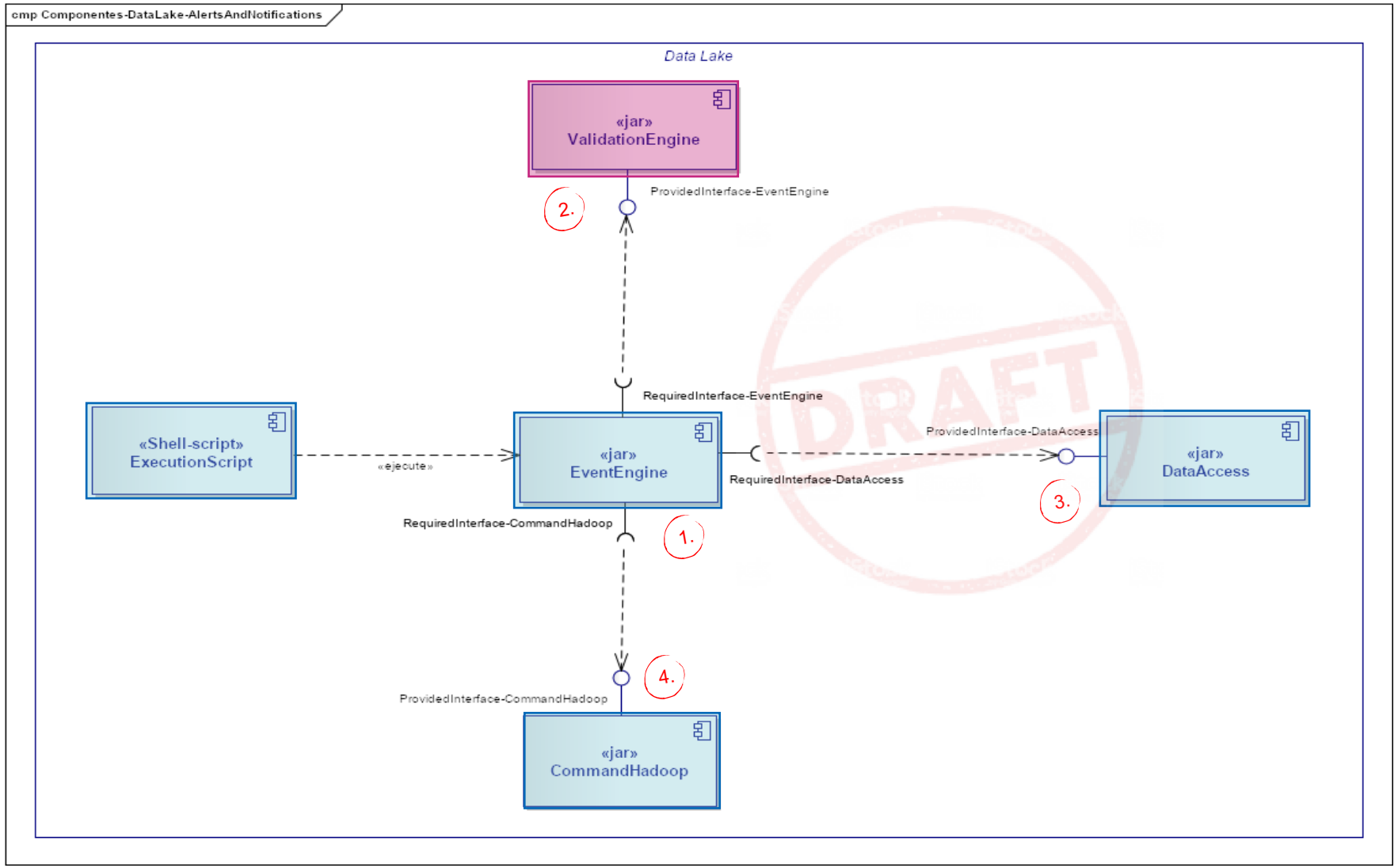


1. Validate File:
2. Move File:
3. Top process data:
4. Apply data quality:
5. create view:

Needs and Solution Proposal (III)

Data Lake - Alerts and Notifications - Component View

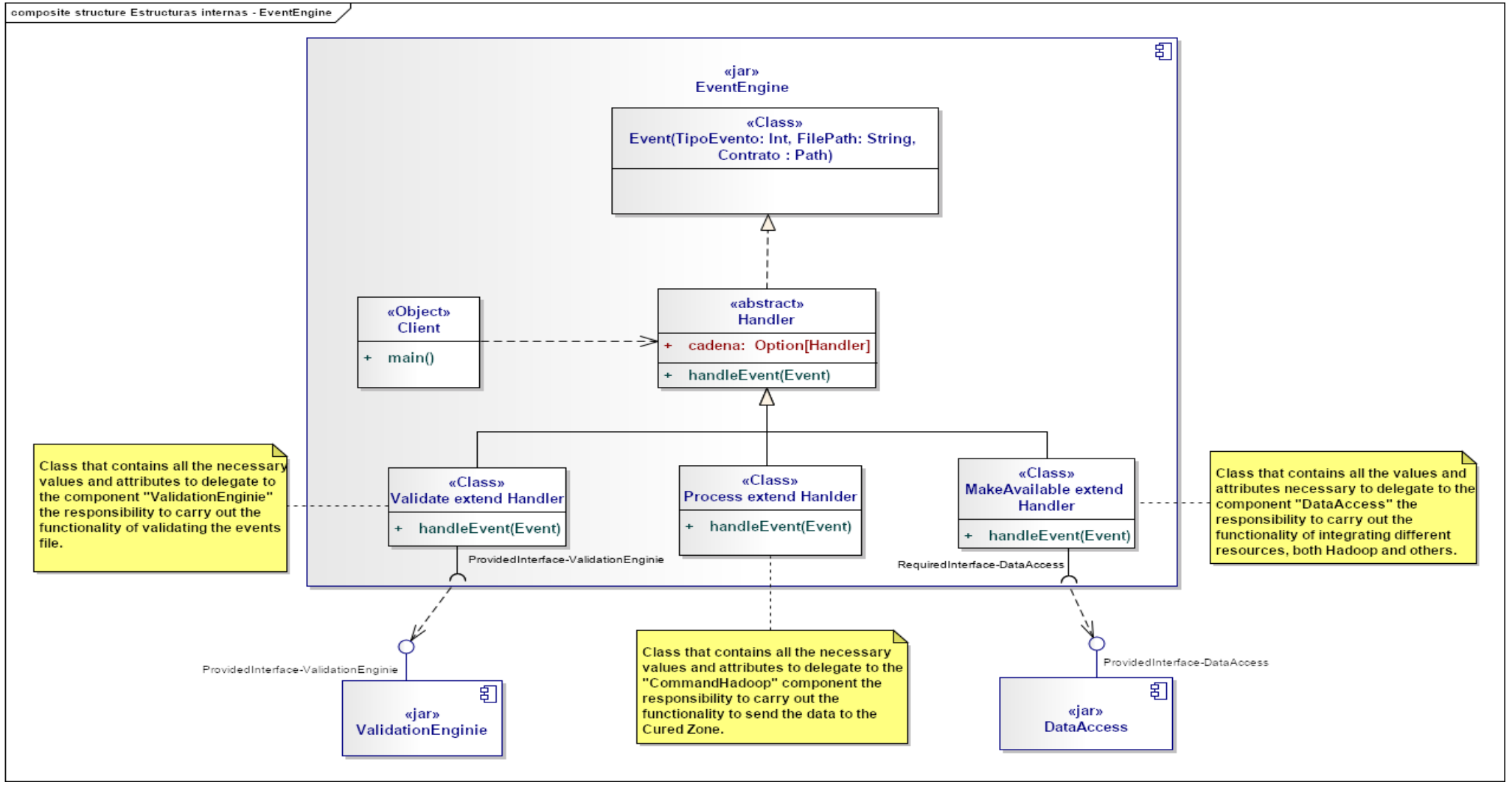
New component
Component to modify
Component not contemplated first phase



1. EventEngine: It is the component that is responsible for providing a common interface to process events, it contains the business logic of different projects that require processing for events
2. validationEngine: It is the component that has the responsibility to perform the functionality of validating the structures of the events, it is contemplated to scale the component for validations in different projects, allowing reuse
3. DataAccess: It is the component that has the responsibility to provide a common interface for access to different resources, both Hadoop and outside the Data Lake
4. commandHadoop: It is the component that is responsible for providing a common interface to execute Hadoop commands

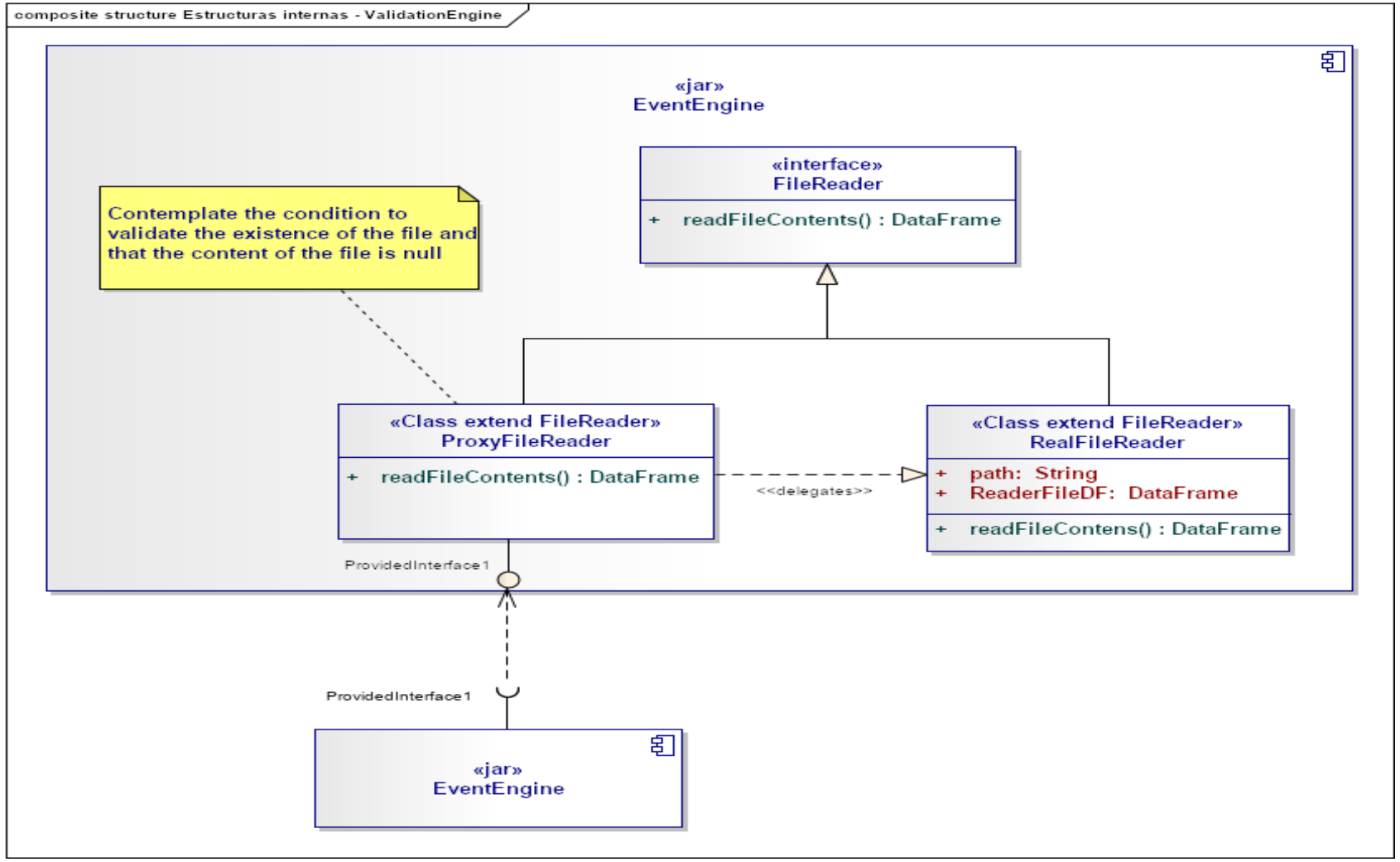
Needs and Solution Proposal (IV)

Data Lake - Alerts and Notifications - Component View II - EventEngine Internal structure



Needs and Solution Proposal (V)

Data Lake - Alerts and Notifications - Component View II - ValidationEngine Internal structure



Needs and Solution Proposal (VI)

Data Lake - Alerts and Notifications - Component View II - CommandHadoop Internal structure

Needs and Solution Proposal (VII)

Data Lake - Alerts and Notifications - Component View II - DataAccess Internal structure

Needs and Solution Proposal (XIII)

Event Hub Architecture – Deploy

Data completeness

Data completeness

Resumen de Componentes de Infraestructura

With the fibnality of meeting the data complitness requirements, we present a generic view of the dimensions, and those we contemplate should be measured, based on near real time projects.

- The quality of the data can refer to the extent of the data (i.e. the values of the data), or the tension of the data (the scheme)
- The most common dimensions of data quality:

Dimensiones	Definiciones
Accesibilidad	Grado de disponibilidad de los datos o recuperación fácil y rápida
Cantidad apropiada de datos	Grado de cual volumen de datos es apropiado para la tarea en cuestión
Creibilidad	Grado en que los datos se consideran verdaderos y creíbles
Completo	Grado en el que no faltan datos y es de suficiente amplitud y profundidad para la tarea en cuestión
Representación consistente	Grado en que los datos se presentan en el mismo formato
Facilidad de manipulación	Grado en que los datos son fáciles de manipular y aplicar a diferentes tareas
Libre de errores	Grado en que los datos son correctos y confiables
Interpretabilidad	Grado en que los datos están en los idiomas, símbolos y unidades apropiados, y las definiciones son claras
Objetividad	Grado en que los datos son imparciales, sin prejuicios e imparciales
Pertinencia	Grado en que los datos son aplicables y útiles para la tarea en cuestión
Reputación	Grado en que los datos son altamente considerados en términos de su fuente y contenido

Data completeness

Data Quality Dimensions

Dimensiones	Definiciones
Seguridad	Grado en que el acceso a los datos se restringe de forma adecuada para mantener su seguridad
Oportunidad	Grado en que los datos están suficientemente actualizados
Comprensibilidad	Grado en que los datos se comprenden fácilmente
Valor añadido	Grado en que los datos son beneficiosos y proporciona ventajas de sus usos
Exactitud	La cercanía de la representación de un fenómeno de la vida real que un valor de datos intenta representar; medido por las funciones de comparación de distancia de edición
Correcto	Se denomina precisión sintáctica, que se refiere a la cercanía de un valor de datos con respecto a un dominio
Predominio	Preocupa la rapidez con que se actualizan los datos; puede medirse por los últimos metadatos actualizados
Volatilidad	Caracteriza la frecuencia con la cual los datos varían en el tiempo; métrica dada por la longitud de los datos de tiempo sigue siendo válida
Consistencia	Se refiere a la violación de las reglas semánticas definidas sobre los elementos de datos y generalmente expresa una restricción de integridad
Facilidad de comprensión	Cuantos datos son claros, sin ambigüedad y fácilmente comprensibles

Infrastructure and Connectivity

Infrastructure and Connectivity (I)

Resumen de Componentes de Infraestructura

Infrastructure component	Description	Applicability (New, Modified, Unmodified)	Application in which the component is located
Pendiente	Pendiente	Pendiente	Pendiente

Volumetría (I)

Case of use -

Characteristic	Detalle

Servicios Generales (I)

Seguridad

El detalle de los lineamientos mencionados están definidos en los documentos:

- Standards for the creation of topics, development and security

Confidencialidad

- Communication through mutual authentication and Kerberos between Data Lake

Auditoria:

- Sending Logs to the Error Topic

Risks and Dependencies (I)

Coexistence

The detail of the mentioned guidelines are defined

- Validate the integration standards with Data Lake, with Data Governance

Risks

El detalle de los lineamientos mencionados están definidos en los documentos:

- Concept tests to integrate Kerberos between Data Lake

Risks and Dependencies (II)

Dependencias

El detalle de los lineamientos mencionados están definidos en los documentos:

- Standard of Security Controls in Data Lake

Standards, Patterns and Decisions of Architecture (I)

Exceptions to Standards and Architecture Patterns

- Pendiene

Standards, Patterns and Decisions of Architecture (II)

Architecture Decisions

Decisión	Premisas	Razonamiento de la Decisión
Pendienyte	Pendiente	• Pendiente

Standards, Patterns and Decisions of Architecture (III)

Outbuildings

Dependencia	Componente externo a la solución que impacta	Fecha estimada en la que se requiere disponible
Not apply	Not apply	Not apply

References and Annex (I)

References

Nombre y version	Fecha	Comentarios	Rol/Departamento