MX2020 Big Data

GCG Mexico – Event Hub/Data Lake Architecture

October 17, 2019

Versión	Fecha	Descripción del Cambio	Autor/Departamento
0.1	17/10/2019	Creación del documento	[Big Data Architecture]

MX2020 Big Data

Table of Contents

Definición del documento

Purpose

The purpose of this document is to graphically represent the evolution of the Big Data Analytics Framework to meet the functional requirements that are:

- Ingest events from the Event Hub / Notifications and Alerts in the Row area
- Perform the Data Quality process
- Perform the data life cycle process
- Perform the process for data exploitation
- Perform the integration process with the visualization tool
- Perform the report generation process
- Perform the Cold storage process (with a 6 month sale)

The systematic qualities that are intended to meet:

- Security
- Modularity
- Reusability
- High availability
- Fault tolerance
- Scalability

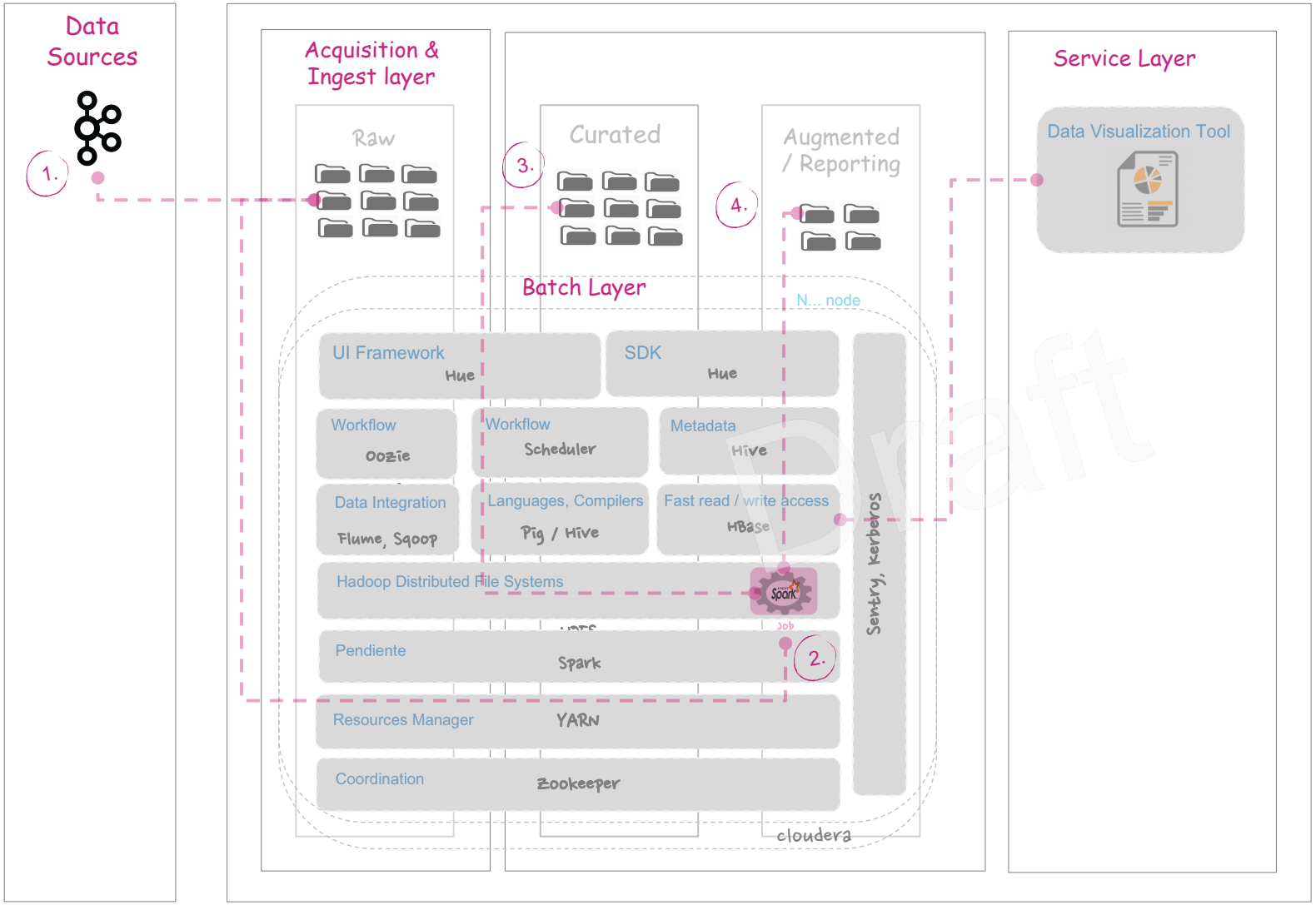
Scope

Se plantea una solución táctica táctica para atender el en una primera fase el Data Quality como Auto Servicio, que permita escalar en una segunda fase los demás componentes del Framework.

Needs and Solution Proposal (I)

Data Lake Global Architecture

New component
Component to modify
Component not covered in scope

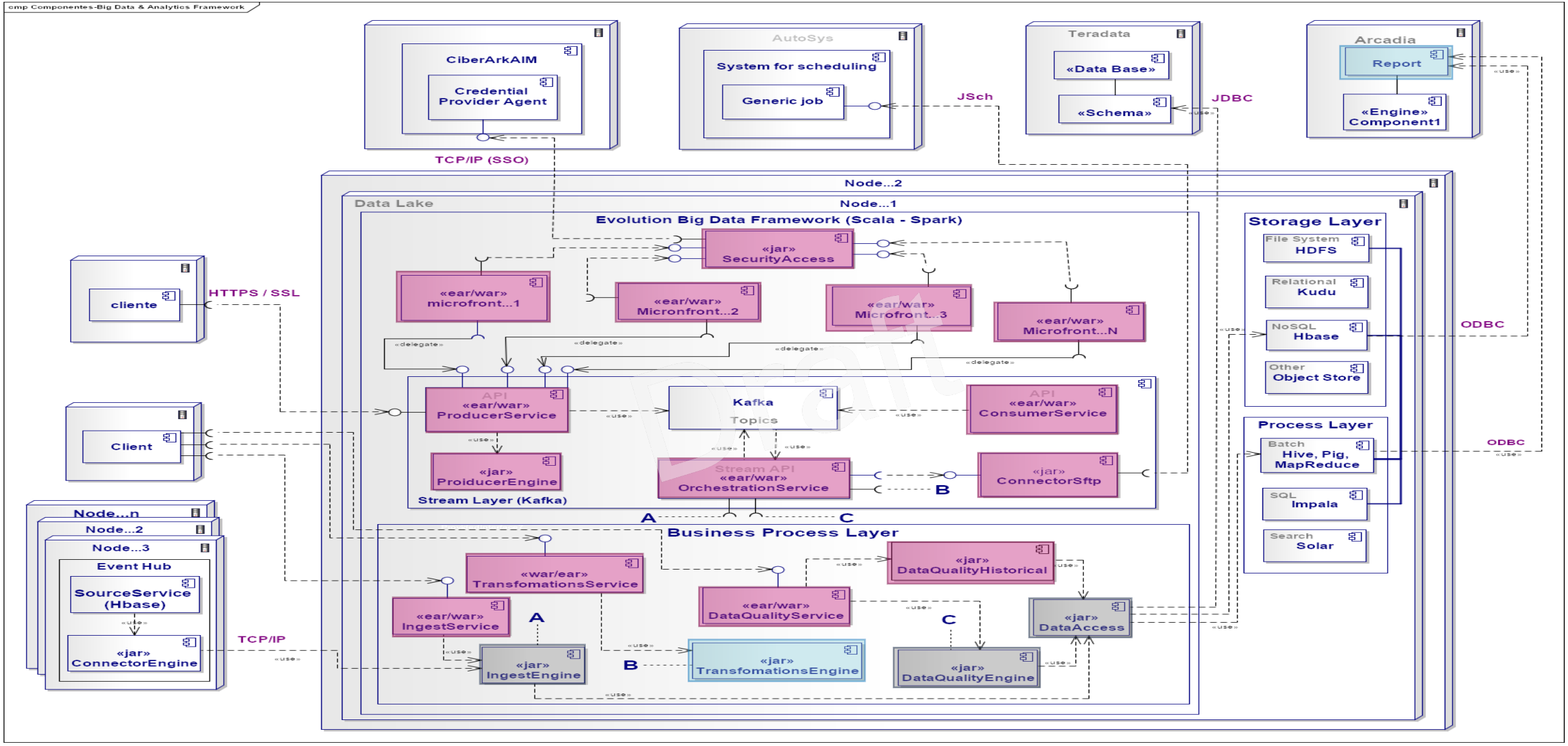


1. Event Hub send a event
2. Spark components perform the data quality process, send to cured area
3. The Spark component performs the data quality process, shipping to cured area
4. The Spark component performs the data quality process, shipping to cured area

Needs and Solution Proposal (I)

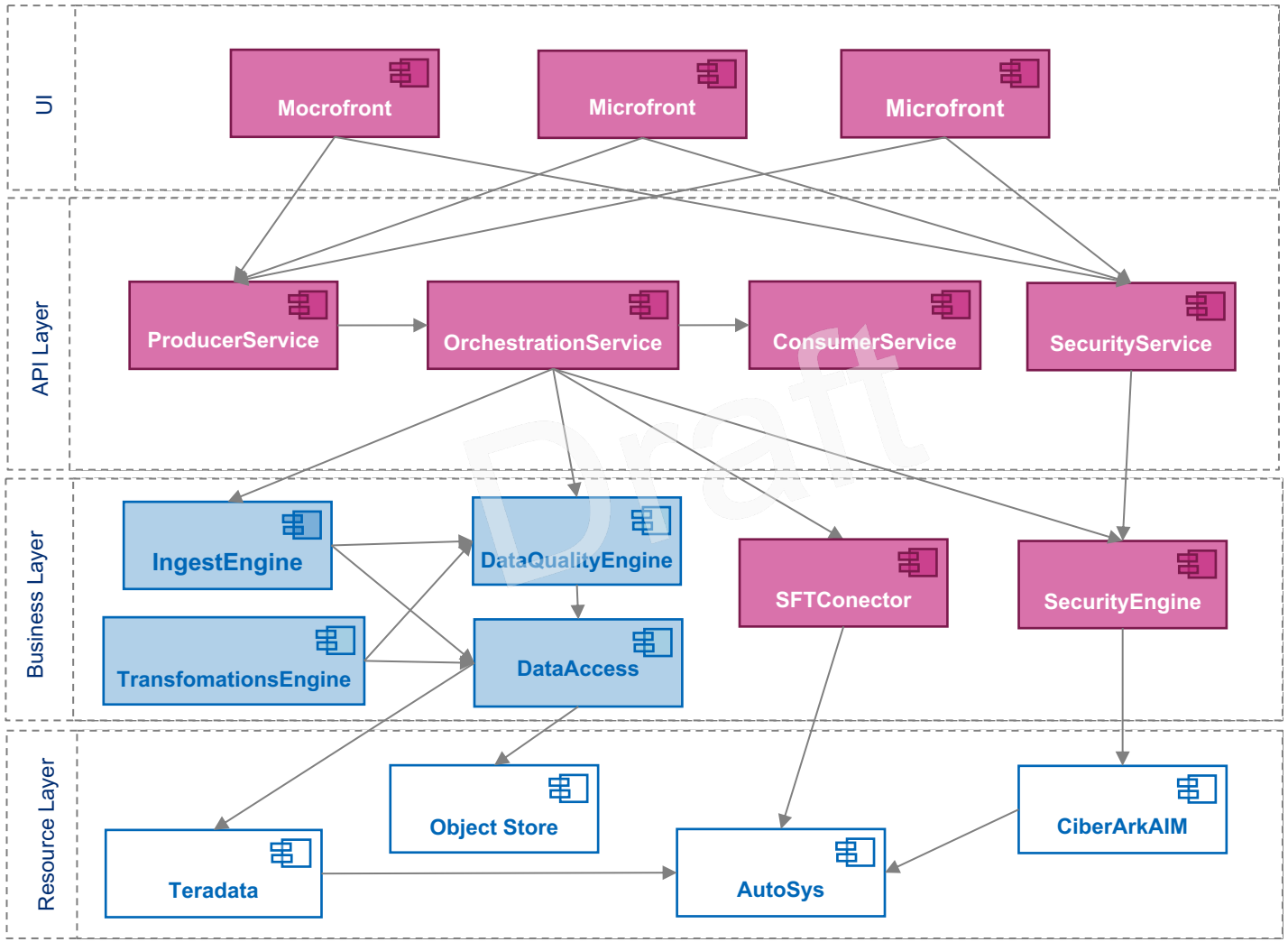
Data Lake Global Architecture – Component Diagram

New component
Component to modify
Component out of reach



Needs and Solution Proposal (II)

Data Lake Architecture – Components II



Needs and Solution Proposal (IV)

Data Lake Architecture – Components IV

It is the component that has the functionality of ingesting allows scaling to different sources, the functionality of obtaining events is incorporated

It is the component that has the business logic for the data quality process, which satisfy the data governance requirements

It is the component that decouples integrations to different resources, such as Teradata, Hive, HBase and contemplates the development for Object Store

It is the component that allows you to perform the transformations to create views in Hive and move in different areas of the data life cycle

Needs and Solution Proposal (IV)

Data Lake Architecture – File of Alerts and messaging

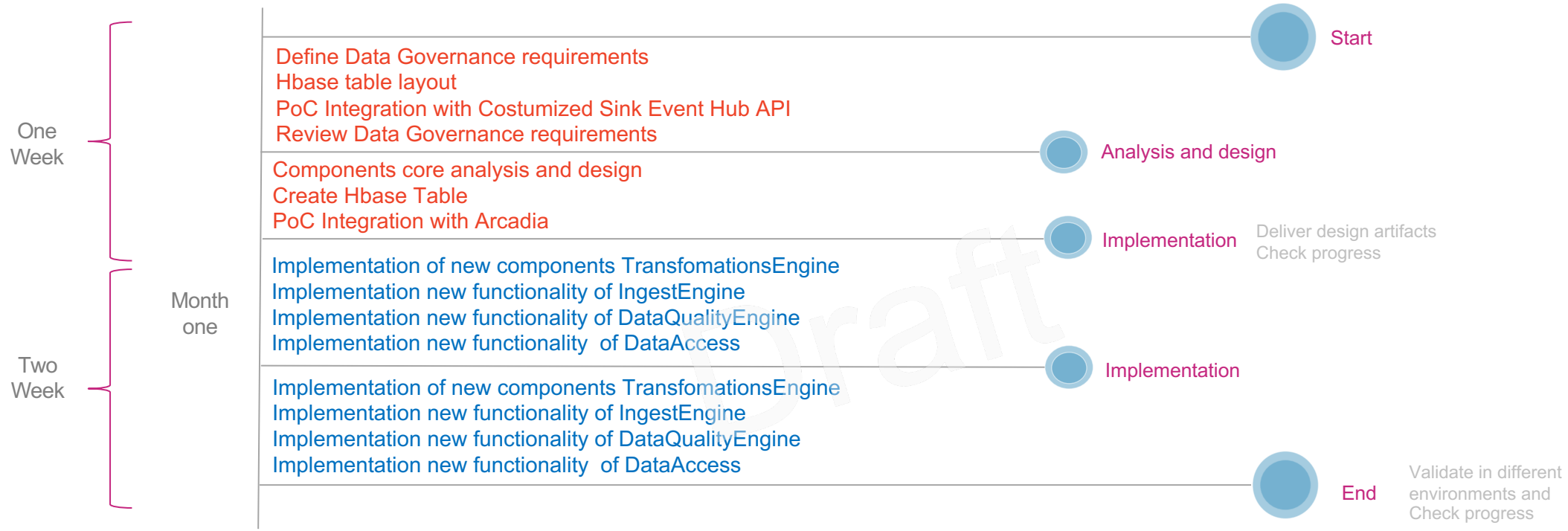
event_id	201677
message_id	54103841810416729134
unique_id	72V91H44M1810X5Q0181C37D0R1M46Q41
message_type	MX010087A201905241354261324
operation_type	24-052019 13:54:26
template_id	A
delivery_massage	{deliveryUUID: , type: , status: , sub_status: , status_description: }
costumer	{ id: 20741677, representative_id: MX570, telephone: , email: armando.antonio.aguilar@gmail.com}
product	{contract_id: 20741677, product_type:AB }
application	{app_id D209F3X78M0152: , app_name: NombreDelMicroServicio }
processing_status	1
status_description	La notificaciòn fallo en el paso 3
processing_time	24-052019 13:54:26

The structure contains fields other than schema and payload, which is the envelope structure used by the JsonConverter with schemas.enable=true (the default).


```
{"schema": {"type": "struct","fields":
  [{"type": "int64", "optional": false, "field": "custumerNumber"},
   {"type": "int32", "optional": false, "field": "BusinessId"},
   {"type": "int32", "optional": false, "field": "countryId"},
   {"type": "int64", "optional": false, "field": "InputMessageId"},
   {"type": "TIMESTAMP", "optional": false, "field": "ProcessDate"},
   {"type": "int32", "optional": false, "field": "communictionType"},
   {"type": "int32", "optional": false, "field": "EventId"},
   {"type": "int32", "optional": false, "field": "ProductId"},
   {"type": "int64", "optional": false, "field": "AccountNumber"},
   {"type": "string", "optional": false, "field": "Amount"},
   {"type": "int64", "optional": false, "field": "AuthorizationNumber"},
   {"type": "string", "optional": false, "field": "MerchantName"},
   {"type": "string", "optional": false, "field": "custumerName"}],
 "optional": false, "name": "ksql.messaging},
 "payload": {
 "custumerNumber": 20741677,
  "BusinessId": 01,
  countryId: MX,
  InputMessageId: MX010087A201905241354261324,
  ProcessDate: 24-052019 13:54:26,
  communictionType: A,
  EventId: 101,
  ProductId: 103,
  AccountNumber: 5634250213811694,
  Amount: +55555555555.55,
  AuthorizationNumber: 12345678,
  MerchantName: Starbucks Portal San Angel,
  custumerName: AGUILR,ARREDONDO/ANTONIO,
  RepId: 99,
  RepName: Antonio Aguilar Arredondo,
  channerlId: EMAIL,
  AlertValue: armando.antonio.aguilar@gmail.com,
  TemplateId: 37,
  communicationcontent: Deposito CUENTA PERFILES M.N. 760 monto $2000.00 el 18/05/19 02:46:00 PM.
En    operaciones con cheque valida tu saldo antes de realizar cualquier operación.,
  ProcessStatus: Error,
  ErrorMessage: Importe inferior al monto mínimo]}
```

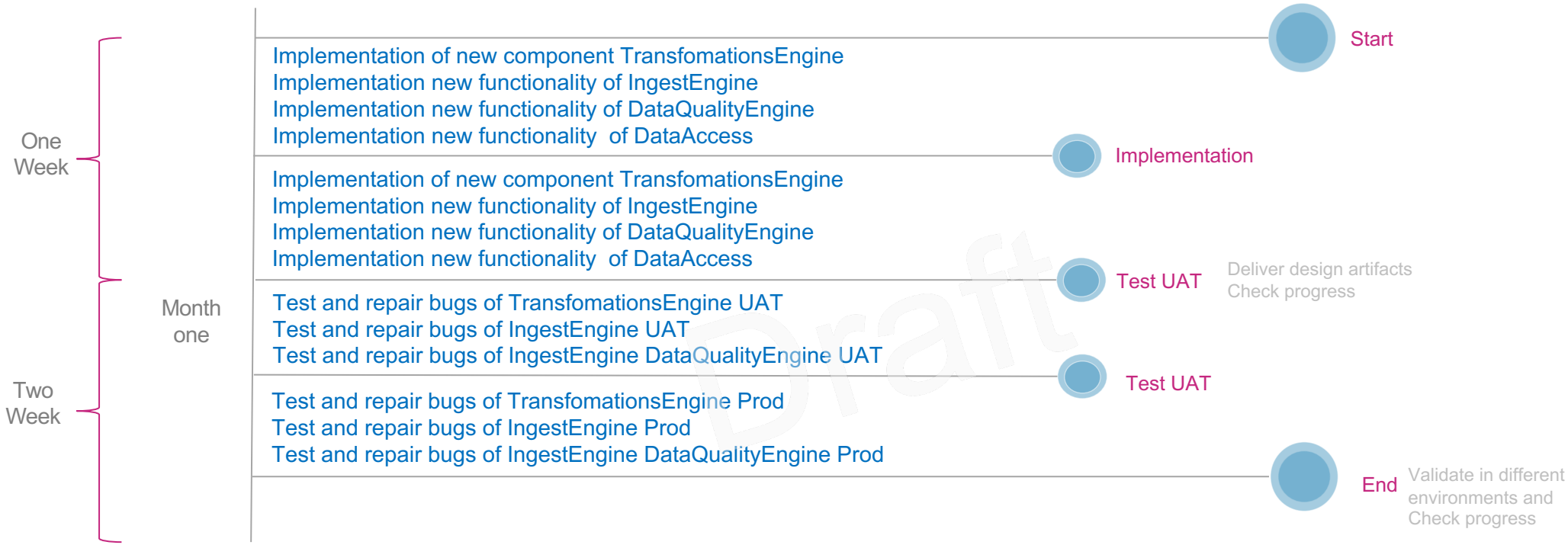
Big data architecture AaaS

Global Architecture – Sprint 1 (Data Lake / Event Hub)



Big data architecture AaaS

Global Architecture – Sprint 2 (Data Lake / Event Hub)

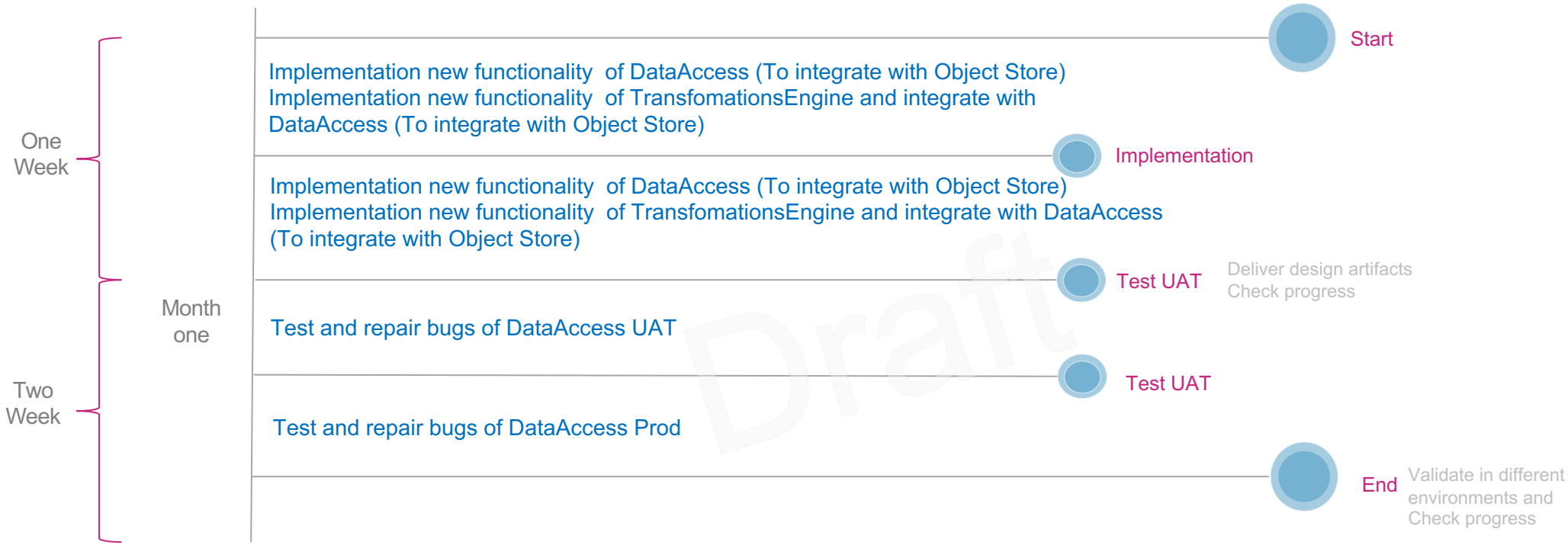


Big data architecture AaaS

Global Architecture – Sprint 3 (Data Lake / Event Hub)

Big data architecture AaaS

Global Architecture – Sprint 4 (Data Lake / Event Hub)



Volumetría (I)

Case of use -

Characteristic	Detalle
number of messages per day	50,000
number of messages per hour	pendiente
event size	pendiente
periodicity	24/7