



Guide to Build a Data Lakehouse

Data-Driven Digital Transformation

Guide to build a data Lakehouse with a business use-case approach



All companies aspire to use more data, but few are able to operationalise the use of data.

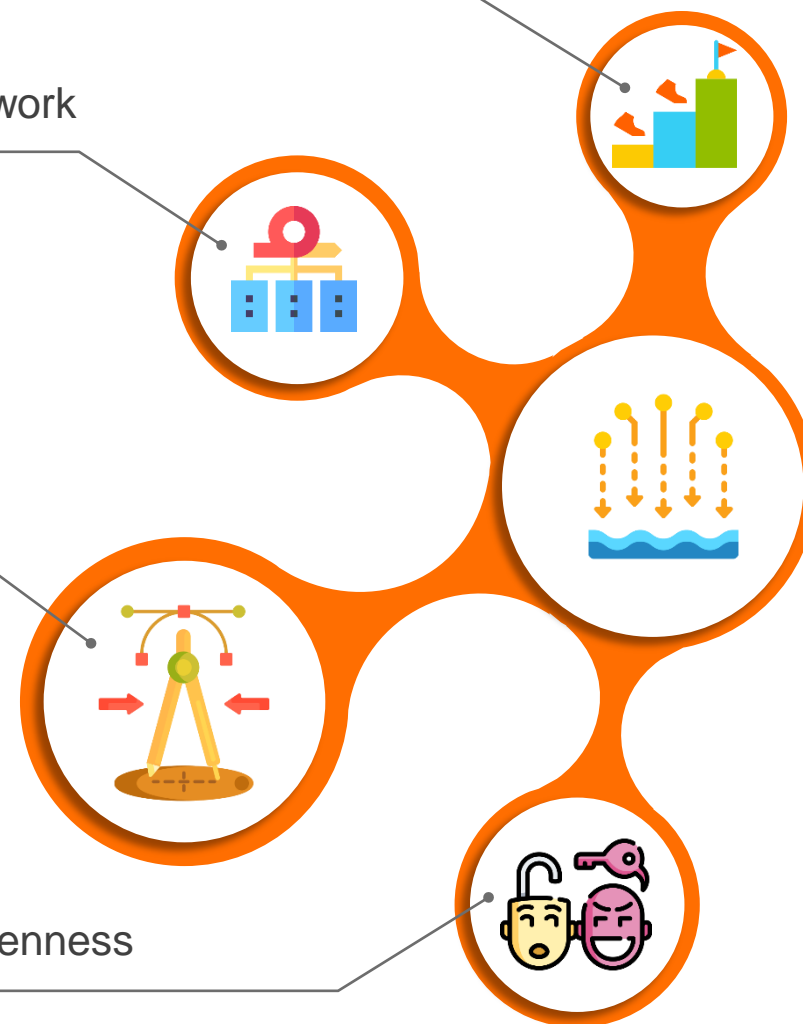
One of the major challenges is putting a broader data strategy into place to make more business-ready information available to analysts and business teams

01 Build and grow data lakehouse – one step at a time

02 Data-driven Ingestion Framework

03 Don't Strive for Perfection

04 Govern with Openness



Build a datahub and grow data lakehouse – one step at time

01 Build and grow data lakehouse – one step at a time

- ✓ Analyse and limit the scope to store required raw data as data hub and from there setting up crossing-functional product teams to develop other layers such conformed & model layer using their own representation specific to their use cases.
- ✓ Start with high-value use cases, focusing on the data needed to deliver those. Expand from them to reach other areas of the business. Apply the lesson learned and drive the adoption to broader business and additional lines of business.
- ✓ A cross-functional delivery team consisting product owners from the business, data engineers and data scientists to productionise the use-cases.
- ✓ Minimize technology risk. This requires focusing on the near-term use cases but keeping an eye on future needs.



Data-drive Ingestion Framework

02 Framework driven data ingestion

- ✓ Create a ingestion a pipeline framework that can store raw data from source systems and aggregate the data from various silos to facilitate faster delivery of business-ready information to analytics teams.
- ✓ Centralised or business-embedded analytic teams can create their own data pipelines, unique to their needs, to drive faster insights and explore new questions.
- ✓ Evolve data delivery pipelines and processes over time to meet the organisation's broader need.



Don't Strive for Perfection

- ✓ Delivery the data available in a timely manner
 - What's the end-to-end data process?
- ✓ Collect the most complete data
 - Is the right data collected?
- ✓ Make sure it is trusted data
 - What's the data quality like?
- ✓ Strive to enable broader data pipelines and more access to data, one step at a time. Start with high-value use cases, focusing on the data needed to deliver those. Expand from there to reach other areas of the business.

03 Don't Strive for Perfection



Govern with Openness

- ✓ Governance should implement with a focus on driving broader use of data. Data security and privacy are important, but this can effectively be balanced with data sharing. A more appropriate and positive way to describe how you govern data is permissible to access. Provide access to those who can properly use the data in effective business situations and don't take an attitude of blocking first.
- ✓ It is important to facilitate use cases where deep governance is less of a requirement and will not be a barrier to adoption. Broader governance policies can be defined and implemented over time as greater data democratisation is facilitated.




04 Govern with Openness



Overview of Data Lakehouse, Data Lake, Data Hub and Data Warehouse

Data Lakehouse	Data Lake	Data Hub	Data Warehouse
----------------	-----------	----------	----------------




Overview of Data Lakehouse, Data Lake, Data Hub and Data Warehouse

	Data Lakehouse	Data Lake	Data Hub	Data Warehouse
Data Ingestion 	<ul style="list-style-type: none">• Load raw data (structured, unstructured & semi-structured) as is• Data stored as object store and transformed in different zones (raw, conformed & curated)• Data governance & data catalogue are applied on different zone	<ul style="list-style-type: none">• Load raw data (structured, unstructured & semi-structured) as is• Data stored as object store and transformed in different zones (raw, conformed & curated)	<ul style="list-style-type: none">• Load raw data (structured, unstructured & semi-structured) as• Raw data persisted in object store	<ul style="list-style-type: none">• Load structured data and transformed to business specific Data mart





Overview of Data Lakehouse, Data Lake, Data Hub and Data Warehouse

	Data Lakehouse	Data Lake	Data Hub	Data Warehouse
Data Ingestion 	<ul style="list-style-type: none"> • Load raw data (structured, unstructured & semi-structured) as is • Data stored as object store and transformed in different zones (raw, conformed & curated) • Data governance & data catalogue are applied on different zone 	<ul style="list-style-type: none"> • Load raw data (structured, unstructured & semi-structured) as is • Data stored as object store and transformed in different zones (raw, conformed & curated) 	<ul style="list-style-type: none"> • Load raw data (structured, unstructured & semi-structured) as • Raw data persisted in object store 	<ul style="list-style-type: none"> • Load structured data and transformed to business specific Data mart
Data Quality 	<ul style="list-style-type: none"> • Data in raw zone not curated • Rest of the zones are highly curated similar to data warehouse 	<ul style="list-style-type: none"> • Data in raw zone not curated • Rest of the zones are highly curated similar to data warehouse 	Highly curated	Highly curated






Overview of Data Lakehouse, Data Lake, Data Hub and Data Warehouse

	Data Lakehouse	Data Lake	Data Hub	Data Warehouse
Data Ingestion 	<ul style="list-style-type: none"> Load raw data (structured, unstructured & semi-structured) as is Data stored as object store and transformed in different zones (raw, conformed & curated) Data governance & data catalogue are applied on different zone 	<ul style="list-style-type: none"> Load raw data (structured, unstructured & semi-structured) as is Data stored as object store and transformed in different zones (raw, conformed & curated) 	<ul style="list-style-type: none"> Load raw data (structured, unstructured & semi-structured) as Raw data persisted in object store 	<ul style="list-style-type: none"> Load structured data and transformed to business specific Data mart
Data Quality 	<ul style="list-style-type: none"> Data in raw zone not curated Rest of the zones are highly curated similar to data warehouse 	<ul style="list-style-type: none"> Data in raw zone not curated Rest of the zones are highly curated similar to data warehouse 	Highly curated	Highly curated
Operational Capabilities 	<ul style="list-style-type: none"> ACID transactions at scale Real-time data 	<ul style="list-style-type: none"> No ACID transactions Other tools used to operationalize the data Real-time data 	<ul style="list-style-type: none"> ACID transactions at scale Real-time data 	<ul style="list-style-type: none"> ACID transactions Near real-time data







Overview of Data Lakehouse, Data Lake, Data Hub and Data Warehouse

	Data Lakehouse	Data Lake	Data Hub	Data Warehouse
Data Ingestion 	<ul style="list-style-type: none"> Load raw data (structured, unstructured & semi-structured) as is Data stored as object store and transformed in different zones (raw, conformed & curated) Data governance & data catalogue are applied on different zone 	<ul style="list-style-type: none"> Load raw data (structured, unstructured & semi-structured) as is Data stored as object store and transformed in different zones (raw, conformed & curated) 	<ul style="list-style-type: none"> Load raw data (structured, unstructured & semi-structured) as Raw data persisted in object store 	<ul style="list-style-type: none"> Load structured data and transformed to business specific Data mart
Data Quality 	<ul style="list-style-type: none"> Data in raw zone not curated Rest of the zones are highly curated similar to data warehouse 	<ul style="list-style-type: none"> Data in raw zone not curated Rest of the zones are highly curated similar to data warehouse 	Highly curated	Highly curated
Operational Capabilities 	<ul style="list-style-type: none"> ACID transactions at scale Real-time data 	<ul style="list-style-type: none"> No ACID transactions Other tools used to operationalize the data Real-time data 	<ul style="list-style-type: none"> ACID transactions at scale Real-time data 	<ul style="list-style-type: none"> ACID transactions Near real-time data
Governance 	<ul style="list-style-type: none"> Granular security controls RBAC at entity level Advanced encryption 	<ul style="list-style-type: none"> Poor data security and governance 	<ul style="list-style-type: none"> Granular security controls RBAC at entity level Advanced encryption 	<ul style="list-style-type: none"> RBAC at entity level

Overview of Data Lakehouse, Data Lake, Data Hub and Data Warehouse

	Data Lakehouse	Data Lake	Data Hub	Data Warehouse
Data Ingestion 	<ul style="list-style-type: none"> Load raw data (structured, unstructured & semi-structured) as is Data stored as object store and transformed in different zones (raw, conformed & curated) Data governance & data catalogue are applied on different zone 	<ul style="list-style-type: none"> Load raw data (structured, unstructured & semi-structured) as is Data stored as object store and transformed in different zones (raw, conformed & curated) 	<ul style="list-style-type: none"> Load raw data (structured, unstructured & semi-structured) as Raw data persisted in object store 	<ul style="list-style-type: none"> Load structured data and transformed to business specific Data mart
Data Quality 	<ul style="list-style-type: none"> Data in raw zone not curated Rest of the zones are highly curated similar to data warehouse 	<ul style="list-style-type: none"> Data in raw zone not curated Rest of the zones are highly curated similar to data warehouse 	Highly curated	Highly curated
Operational Capabilities 	<ul style="list-style-type: none"> ACID transactions at scale Real-time data 	<ul style="list-style-type: none"> No ACID transactions Other tools used to operationalize the data Real-time data 	<ul style="list-style-type: none"> ACID transactions at scale Real-time data 	<ul style="list-style-type: none"> ACID transactions Near real-time data
Governance 	<ul style="list-style-type: none"> Granular security controls RBAC at entity level Advanced encryption 	<ul style="list-style-type: none"> Poor data security and governance 	<ul style="list-style-type: none"> Granular security controls RBAC at entity level Advanced encryption 	<ul style="list-style-type: none"> RBAC at entity level
Scalability 	<ul style="list-style-type: none"> Petabyte scalability Ideal for low-cost storage Cost can grow exponential without archiving policies Must need hot and cold storage option 	<ul style="list-style-type: none"> Petabyte scalability Ideal for low-cost storage 	<ul style="list-style-type: none"> Petabyte scalability Higher cost due to indexing overhead for some implementations 	<ul style="list-style-type: none"> Only performs as well as the slowest federate, and is impacted by system load or issues in any federate

Overview of Data Lakehouse, Data Lake, Data Hub and Data Warehouse

	Data Lakehouse	Data Lake	Data Hub	Data Warehouse
Data Ingestion 	<ul style="list-style-type: none"> Load raw data (structured, unstructured & semi-structured) as is Data stored as object store and transformed in different zones (raw, conformed & curated) Data governance & data catalogue are applied on different zone 	<ul style="list-style-type: none"> Load raw data (structured, unstructured & semi-structured) as is Data stored as object store and transformed in different zones (raw, conformed & curated) 	<ul style="list-style-type: none"> Load raw data (structured, unstructured & semi-structured) as Raw data persisted in object store 	<ul style="list-style-type: none"> Load structured data and transformed to business specific Data mart
Data Quality 	<ul style="list-style-type: none"> Data in raw zone not curated Rest of the zones are highly curated similar to data warehouse 	<ul style="list-style-type: none"> Data in raw zone not curated Rest of the zones are highly curated similar to data warehouse 	Highly curated	Highly curated
Operational Capabilities 	<ul style="list-style-type: none"> ACID transactions at scale Real-time data 	<ul style="list-style-type: none"> No ACID transactions Other tools used to operationalize the data Real-time data 	<ul style="list-style-type: none"> ACID transactions at scale Real-time data 	<ul style="list-style-type: none"> ACID transactions Near real-time data
Governance 	<ul style="list-style-type: none"> Granular security controls RBAC at entity level Advanced encryption 	<ul style="list-style-type: none"> Poor data security and governance 	<ul style="list-style-type: none"> Granular security controls RBAC at entity level Advanced encryption 	<ul style="list-style-type: none"> RBAC at entity level
Scalability 	<ul style="list-style-type: none"> Petabyte scalability Ideal for low-cost storage Cost can grow exponential without archiving policies Must need hot and cold storage option 	Petabyte scalability Ideal for low-cost storage	<ul style="list-style-type: none"> Petabyte scalability Higher cost due to indexing overhead for some implementations 	<ul style="list-style-type: none"> Only performs as well as the slowest federate, and is impacted by system load or issues in any federate
Users 	<ul style="list-style-type: none"> Data scientists, business users, analysts 	<ul style="list-style-type: none"> Data scientists, business users, analysts 	<ul style="list-style-type: none"> Diverse business users 	<ul style="list-style-type: none"> Business Analysts

Put it to test

Let's illustrate with an example :

A utility company has taken initiative on a data-driven digital transformation to use data to guide transition to new business model. The company doesn't have a view about their customers segmentations and utilities assets nor consolidated view of their customers across different line-of-business such as energy, gas and renewable energy.

Before they embark on data-driven digital transformation they are cognisant of below:

- Committing to zero carbon emissions by adopting clean energy
- How do you transform your business model to exploit the opportunities of big data without risking current revenue and disturbing current BAU process?
- What is the best strategy between the current data estate and the new?

Apply Product Thinking for data-driven digital transformation

The common denominator for every digital transformation is data

Build one data product at a time.

It starts with the following initial business use case:

- Committing to clean energy - Identify customers across different line-of-business such as energy, gas and renewable energy; business goal of converting customers from non-renewable to clean energy
- Cross-Selling and Upselling other products to the customer based on the products they already have; business goal is to increase the conversion rate and decrease customer retention
- Utility Maintenance Alerts – Identify utility assets and preventive and predictive maintenance; business goal to reduce OPEX on assets

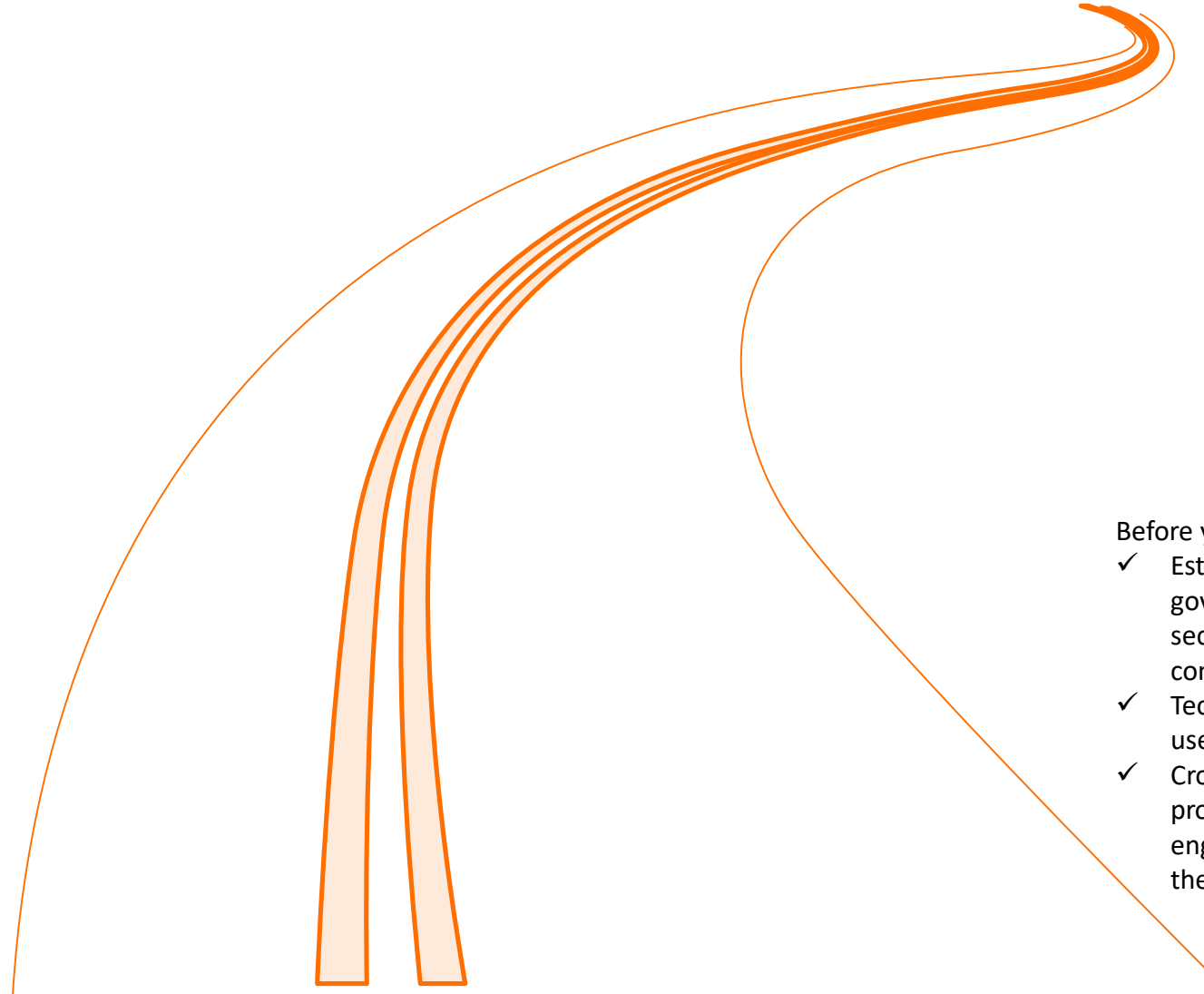
Business use-case Roadmap

- ✓ Start small, pick a business use-case which is easy to implement and helps to establish foundation of data lakehouse platform



Business use-case Roadmap

- ✓ Start small, pick a business use-case which is easy to implement and helps to establish foundation of data lakehouse platform

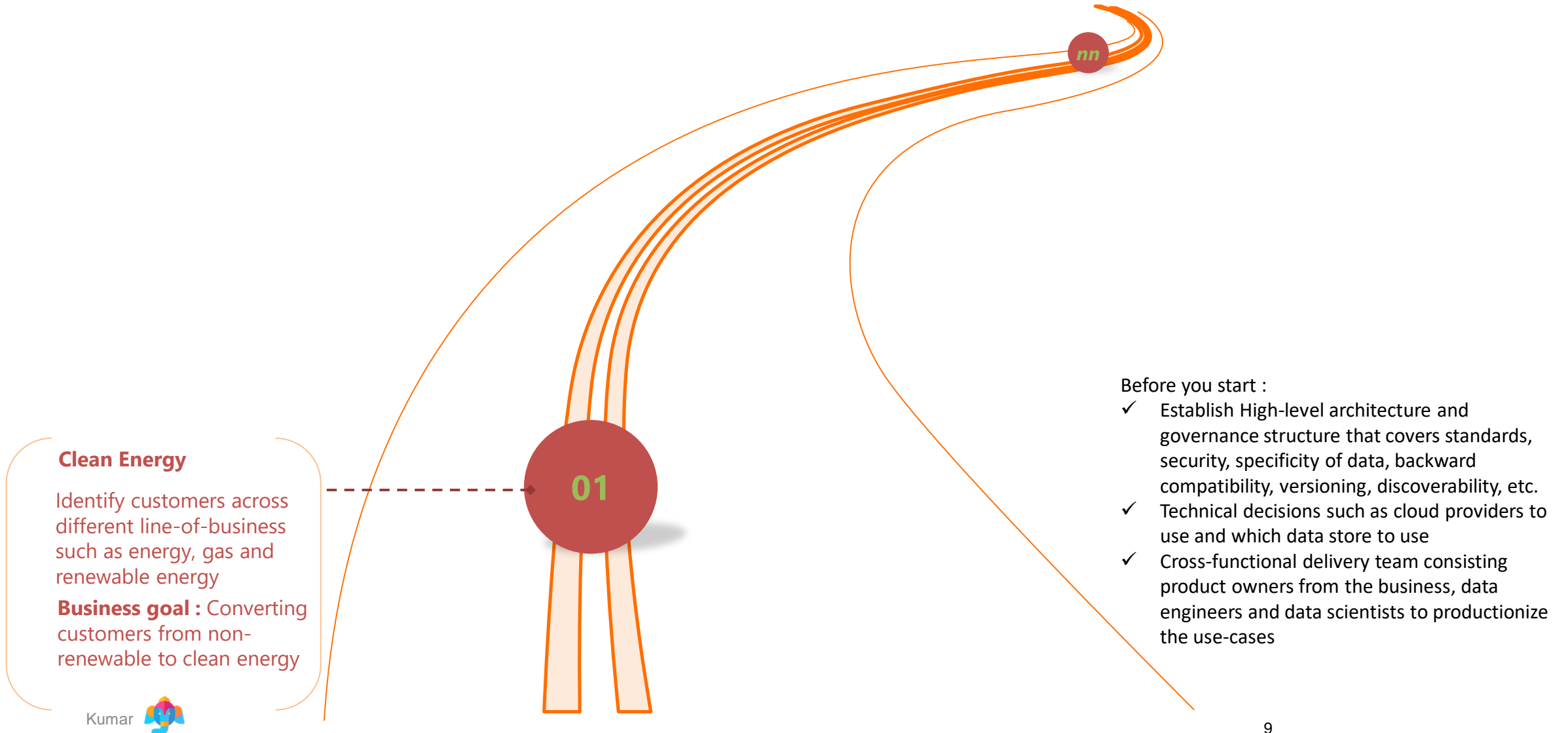


Before you start :

- ✓ Establish High-level architecture and governance structure that covers standards, security, specificity of data, backward compatibility, versioning, discoverability, etc.
- ✓ Technical decisions such as cloud providers to use and which data store to use
- ✓ Cross-functional delivery team consisting product owners from the business, data engineers and data scientists to productionize the use-cases

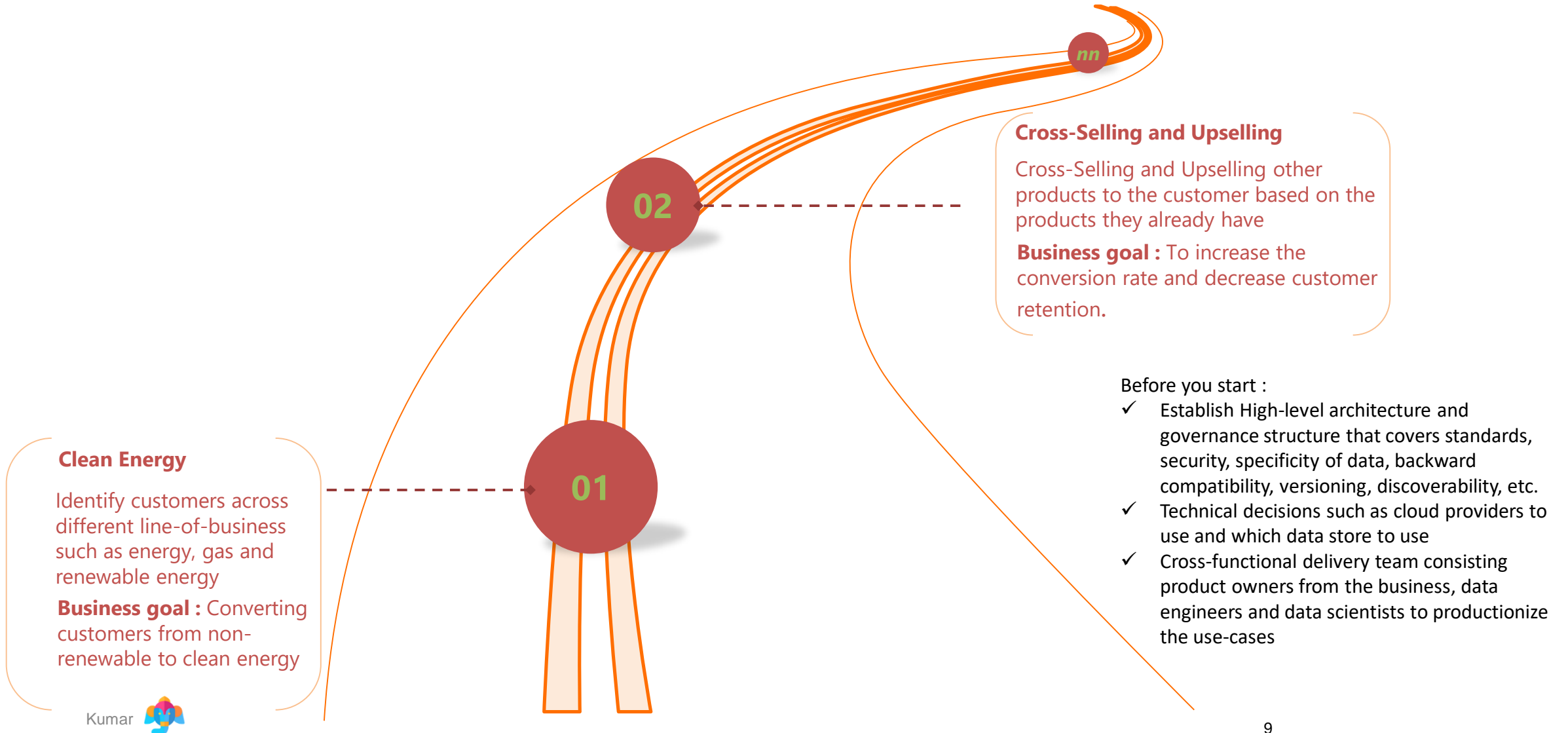
Business use-case Roadmap

- ✓ Start small, pick a business use-case which is easy to implement and helps to establish foundation of data lakehouse platform



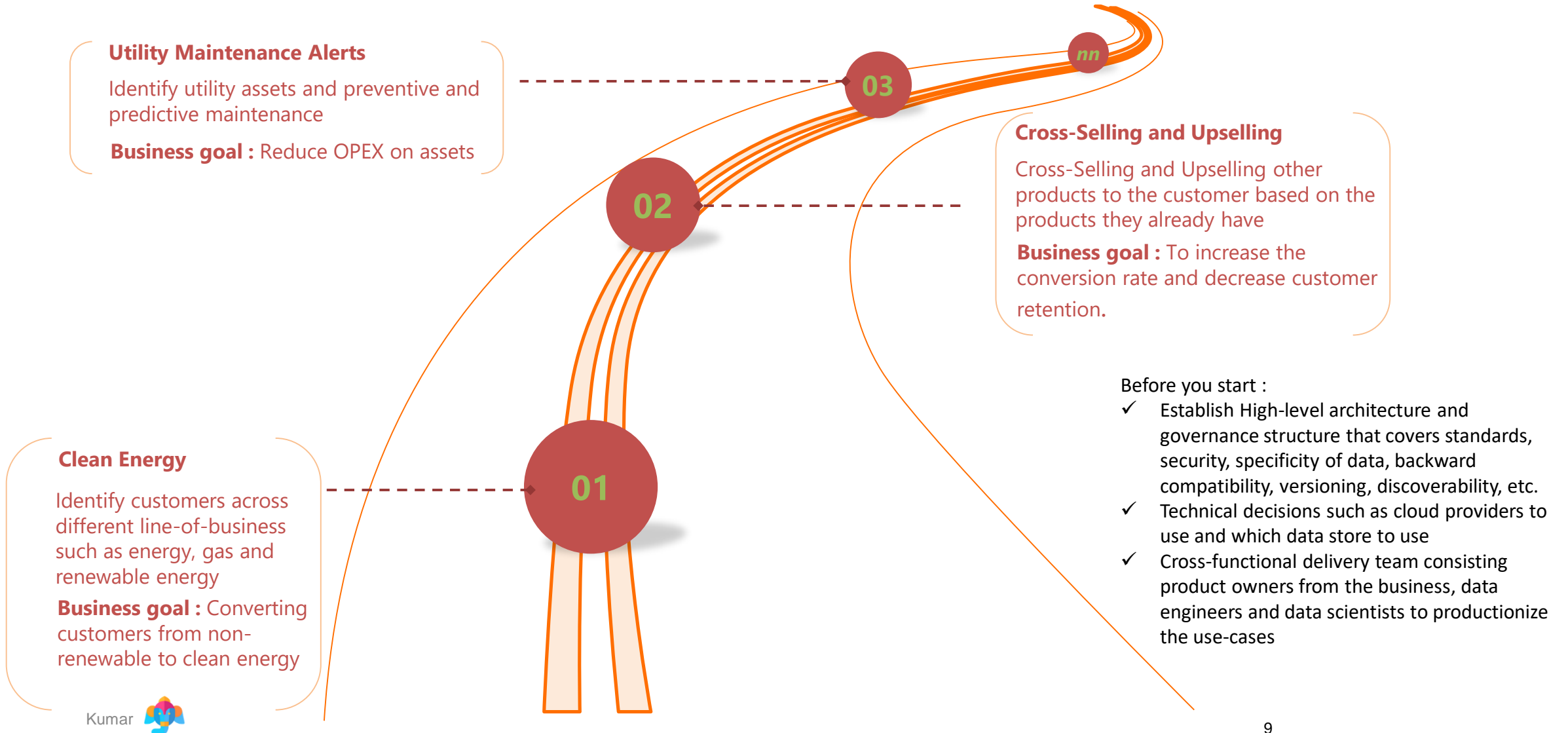
Business use-case Roadmap

- ✓ Start small, pick a business use-case which is easy to implement and helps to establish foundation of data lakehouse platform



Business use-case Roadmap

- ✓ Start small, pick a business use-case which is easy to implement and helps to establish foundation of data lakehouse platform



Walk through Implementation of Data Lakehouse for each Business use-case

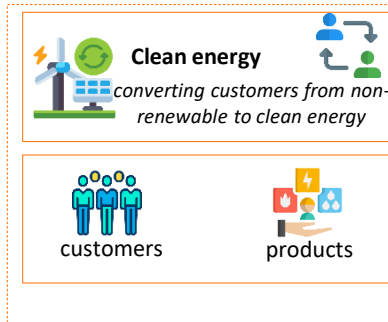
Walk through Implementation of Data Lakehouse for each Business use-case

1

Use-case : Clean energy

1. Ingest raw customer and product into data lakehouse.
2. Transform to represent in a DataMart (fact & dimension) in the consumption layer
3. Next step, build a ML model to identify customers and covert them to clean energy products.

Step-1

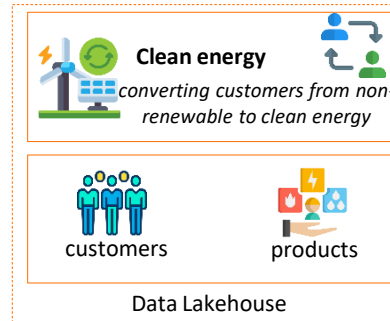


Walk through Implementation of Data Lakehouse for each Business use-case

1 Use-case : Clean energy

1. Ingest raw customer and product into data lakehouse.
2. Transform to represent in a DataMart (fact & dimension) in the consumption layer
3. Next step, build a ML model to identify customers and covert them to clean energy products.

Step-1

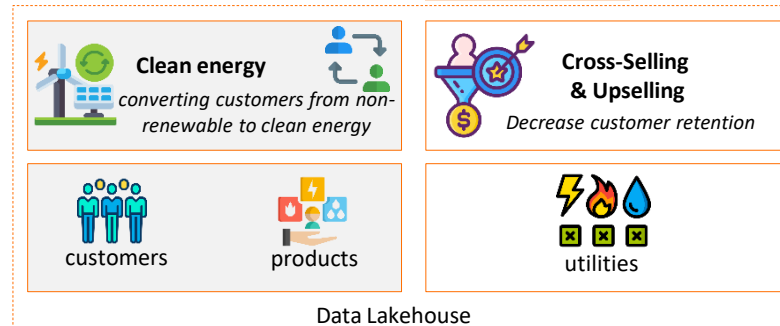


2 Use-case : Cross-Selling and Upselling

Now that clean energy business use-case data product is in production with a good conversion rate; the team can focus on upselling use-case.

1. Ingest data for other products such as gas and water to the existing raw layer
2. Transform the data to build single view of customer in consumption layer
3. Next step, build Next Best Product recommendation engine/ML model for Cross-Selling and Upselling

Step-2

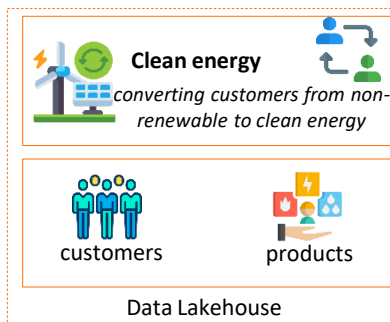


Walk through Implementation of Data Lakehouse for each Business use-case

1 Use-case : Clean energy

1. Ingest raw customer and product into data lakehouse.
2. Transform to represent in a DataMart (fact & dimension) in the consumption layer
3. Next step, build a ML model to identify customers and covert them to clean energy products.

Step-1

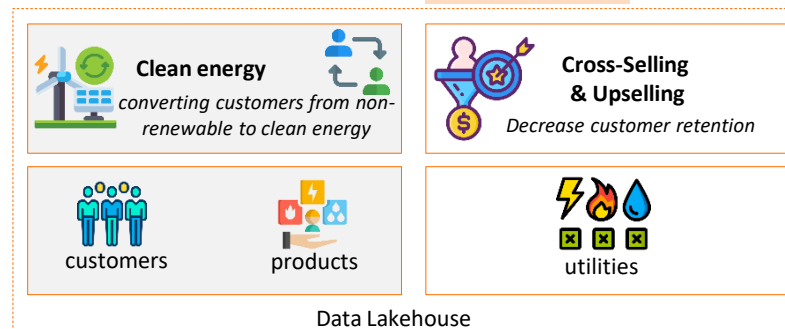


2 Use-case : Cross-Selling and Upselling

Now that clean energy business use-case data product is in production with a good conversion rate; the team can focus on upselling use-case.

1. Ingest data for other products such as gas and water to the existing raw layer
2. Transform the data to build single view of customer in consumption layer
3. Next step, build Next Best Product recommendation engine/ML model for Cross-Selling and Upselling

Step-2



3 Use-case : Utility Assets Alerts (preventive & predictive)

Now that both customer related data products is in production the team can focus on reducing OPEX utility assets.

1. Ingest the data related to different assets and its maintenance history
2. build preventive and predictive maintenance ML models to alert on utility assets.

Step-3

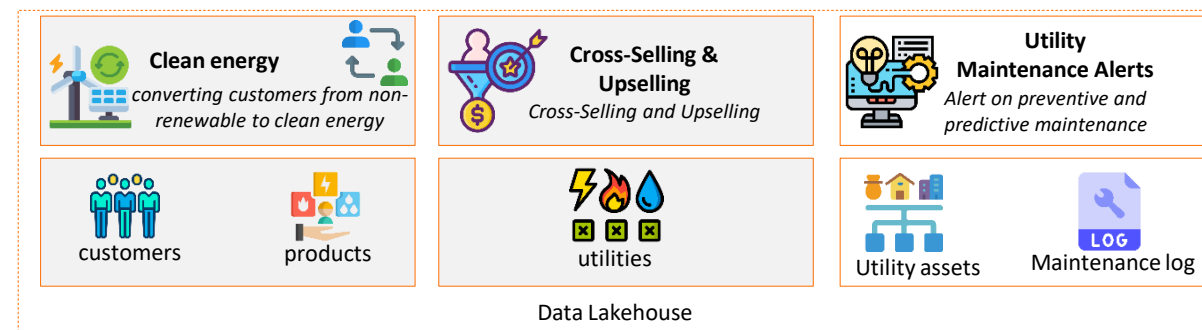


Illustration of Data flow

Illustration of Data flow

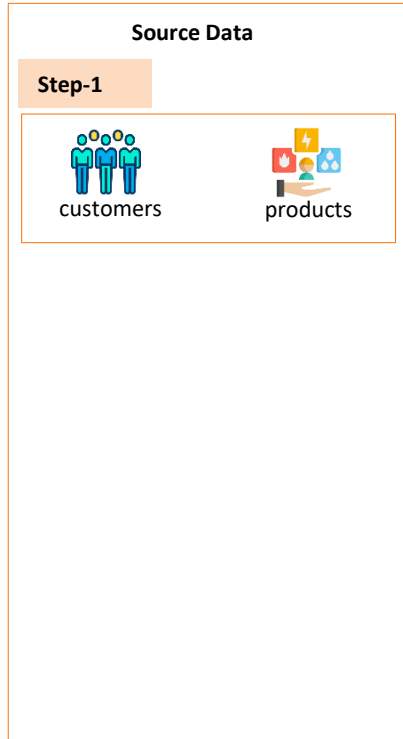


Illustration of Data flow

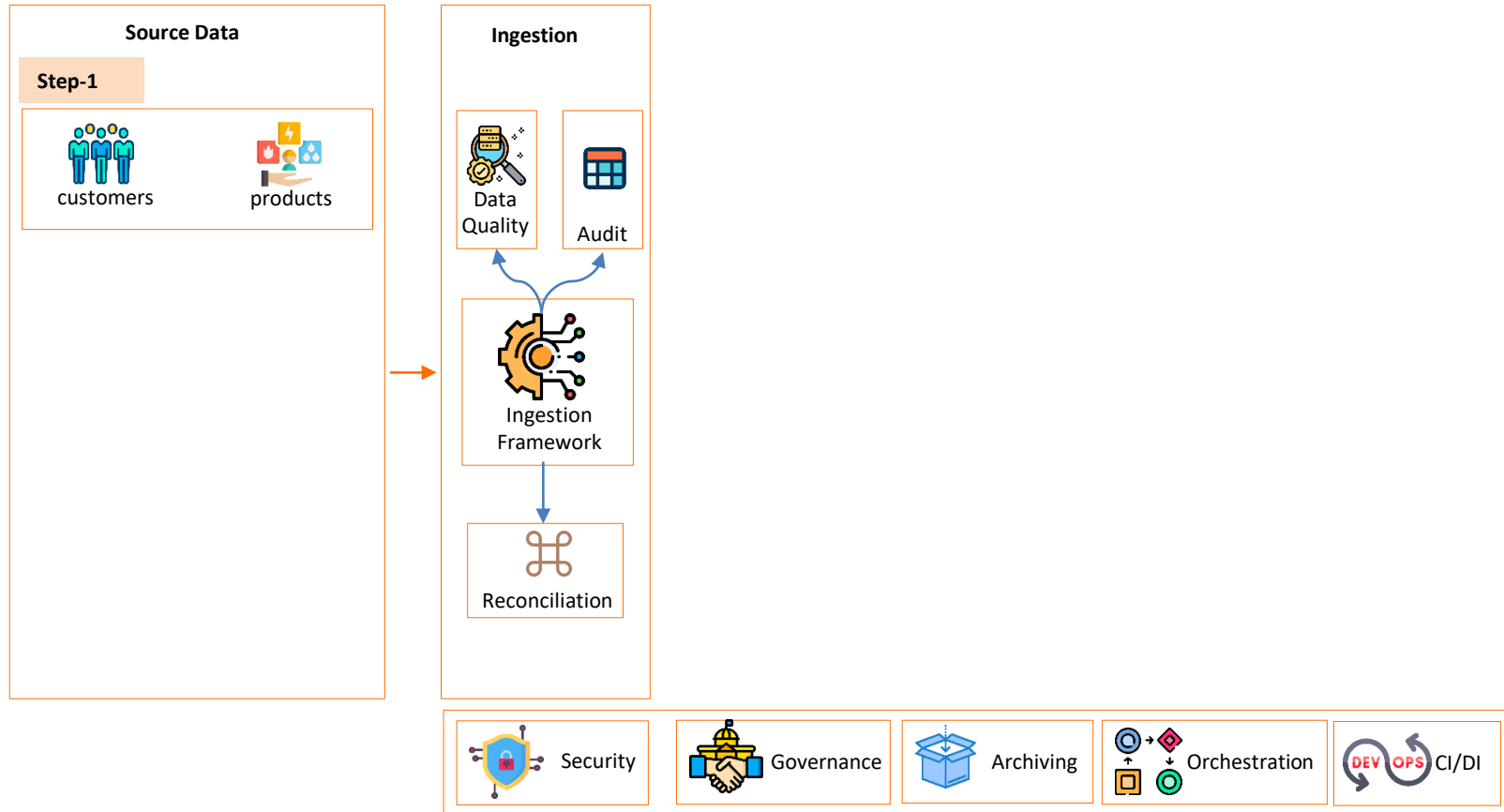


Illustration of Data flow

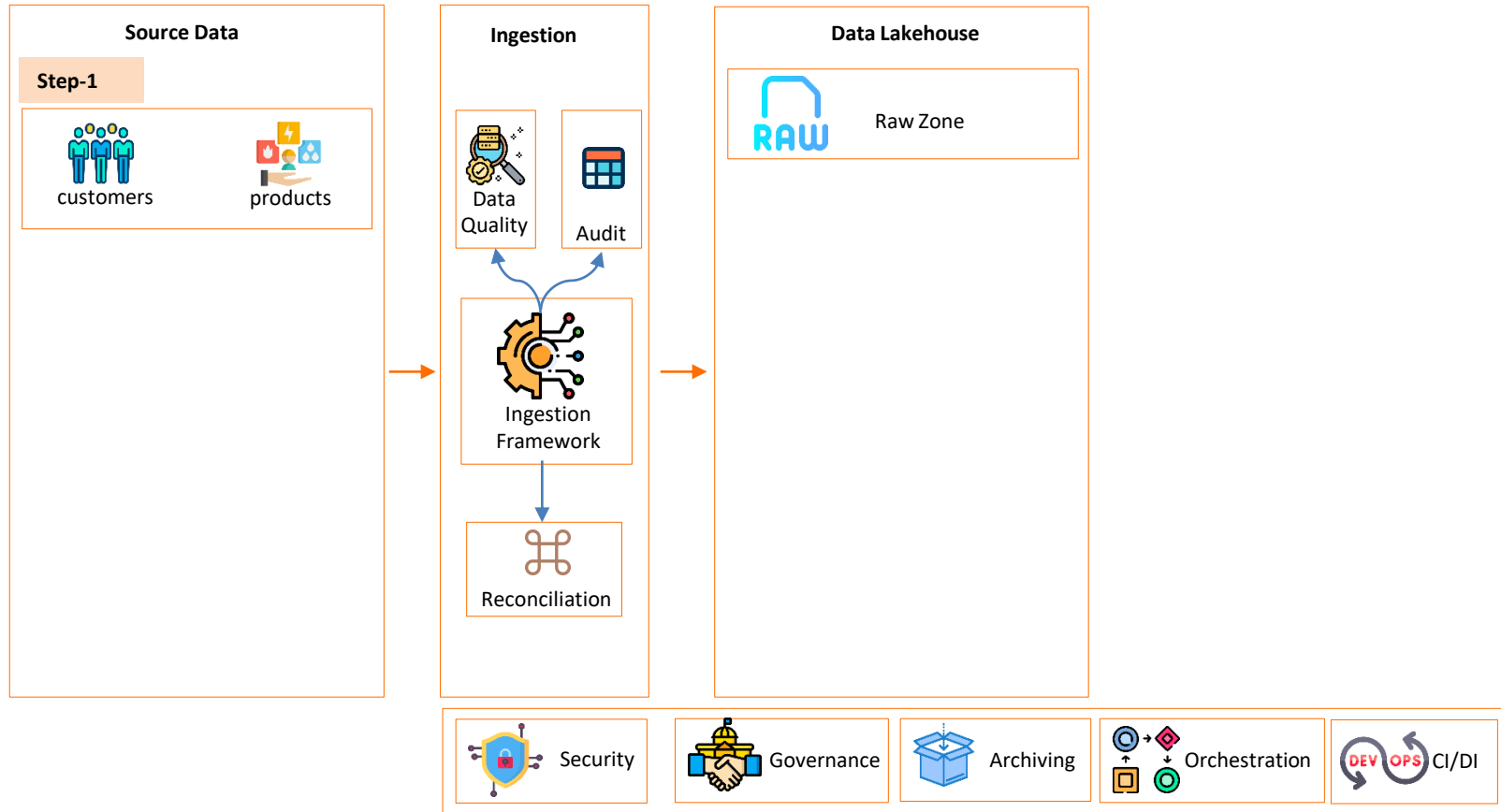


Illustration of Data flow

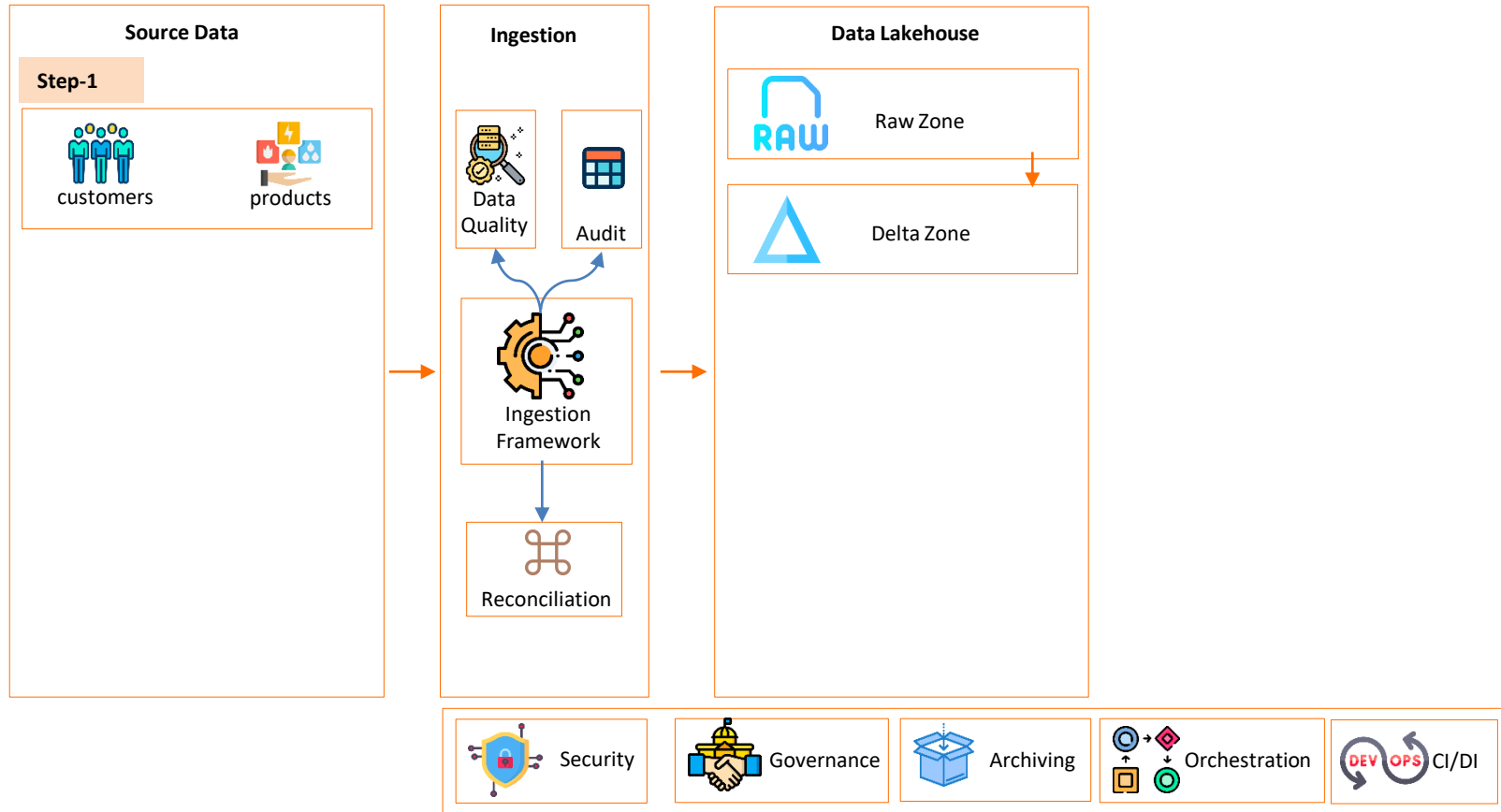


Illustration of Data flow

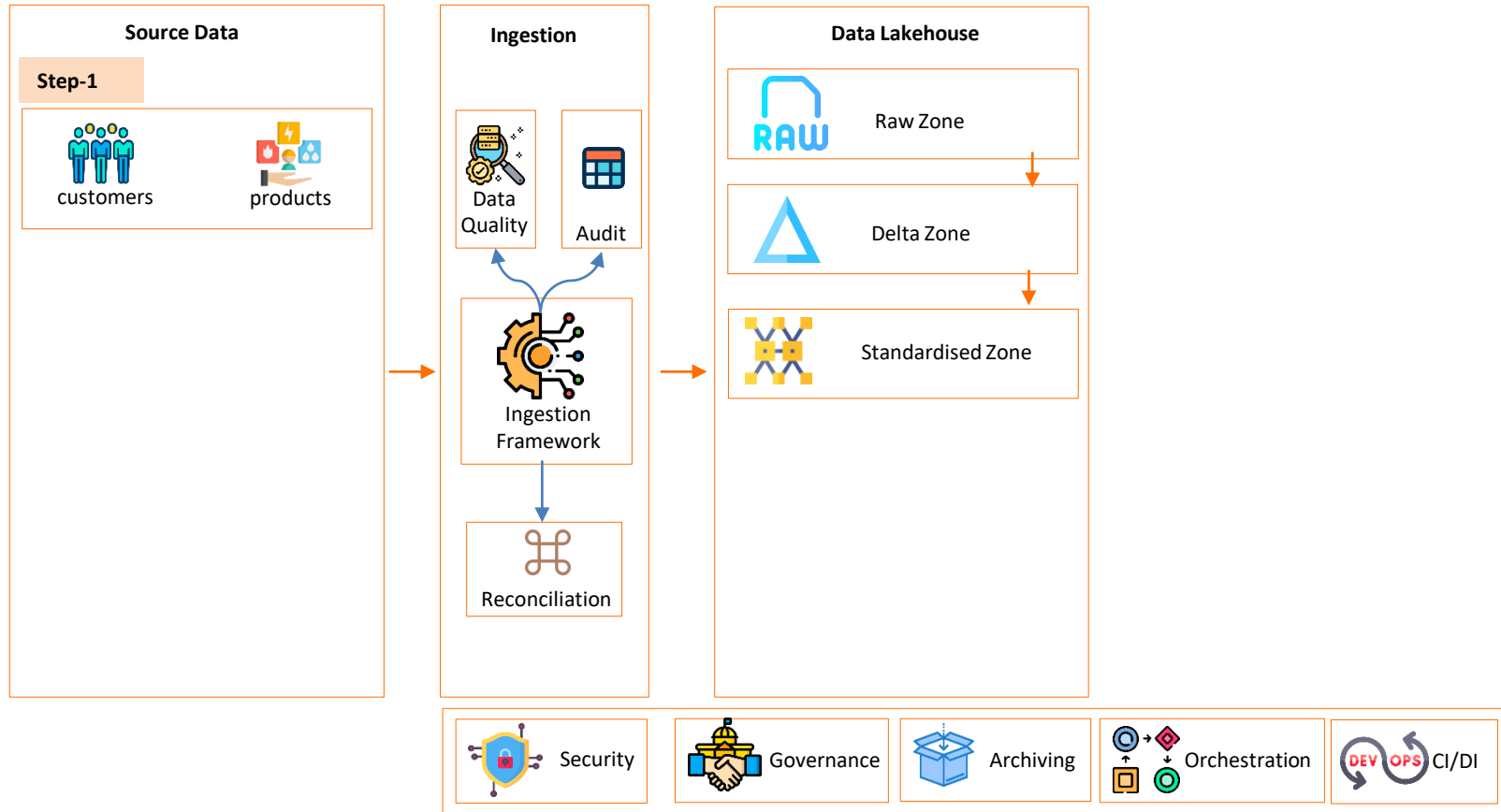


Illustration of Data flow

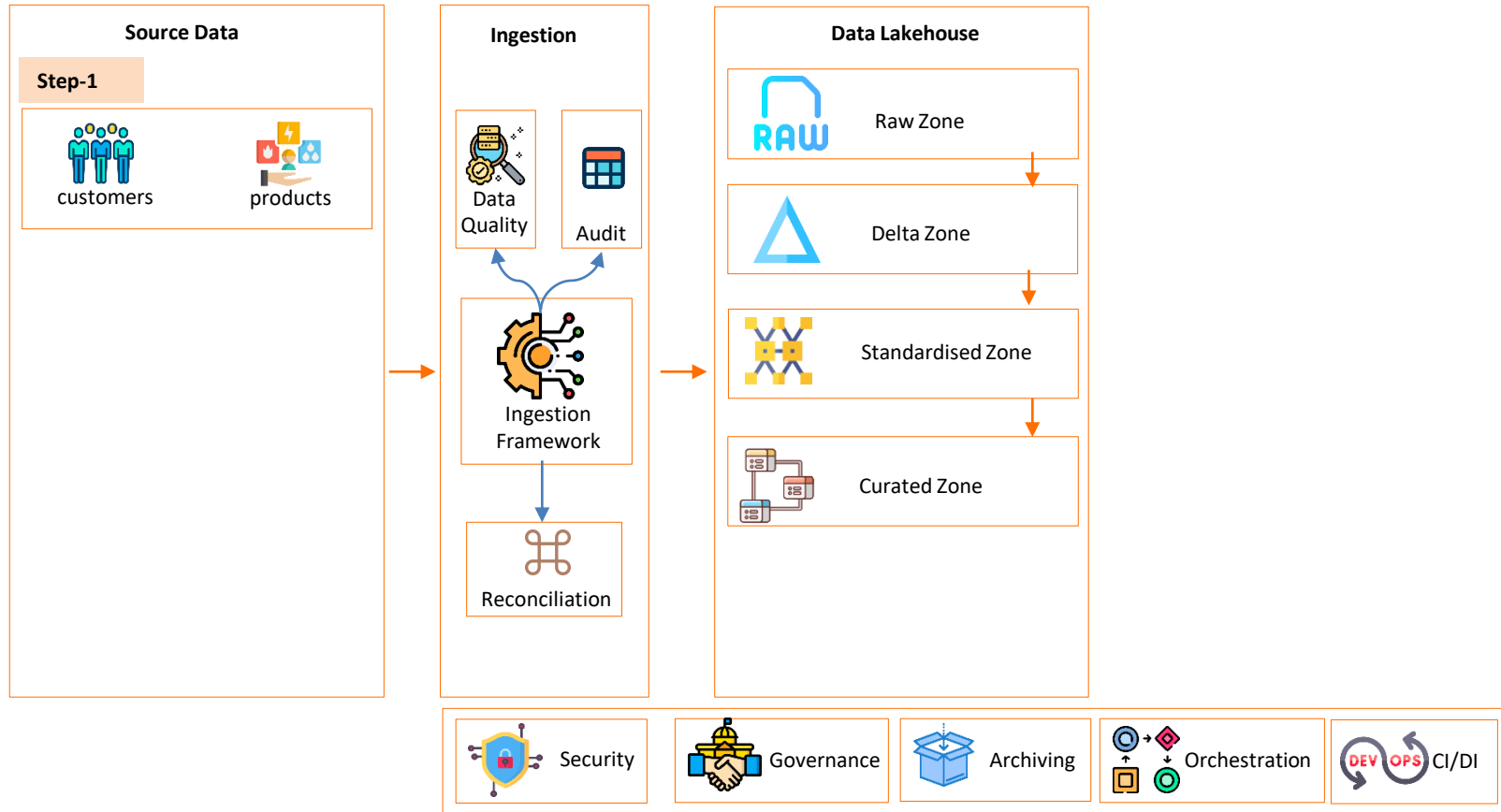


Illustration of Data flow

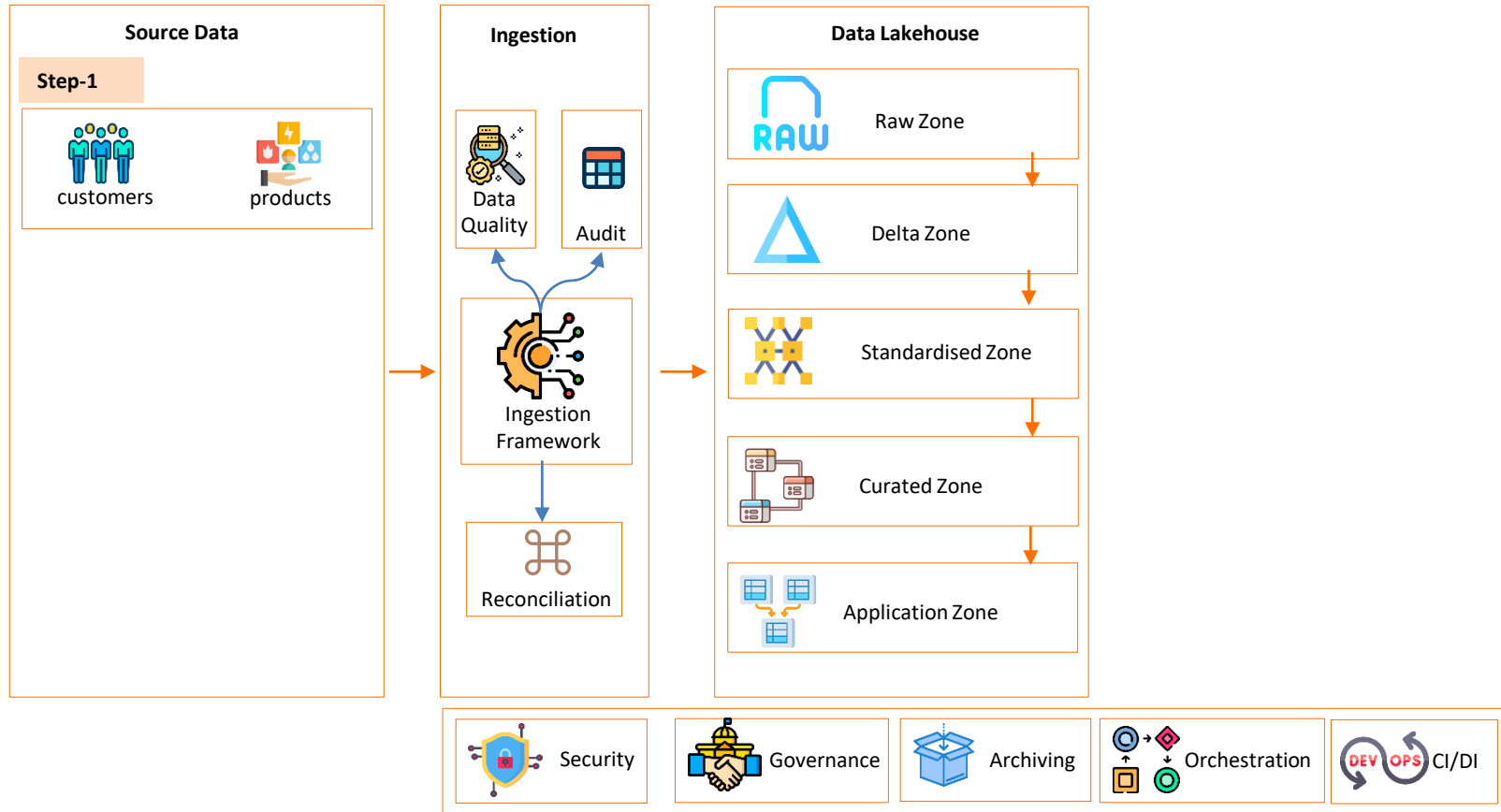


Illustration of Data flow

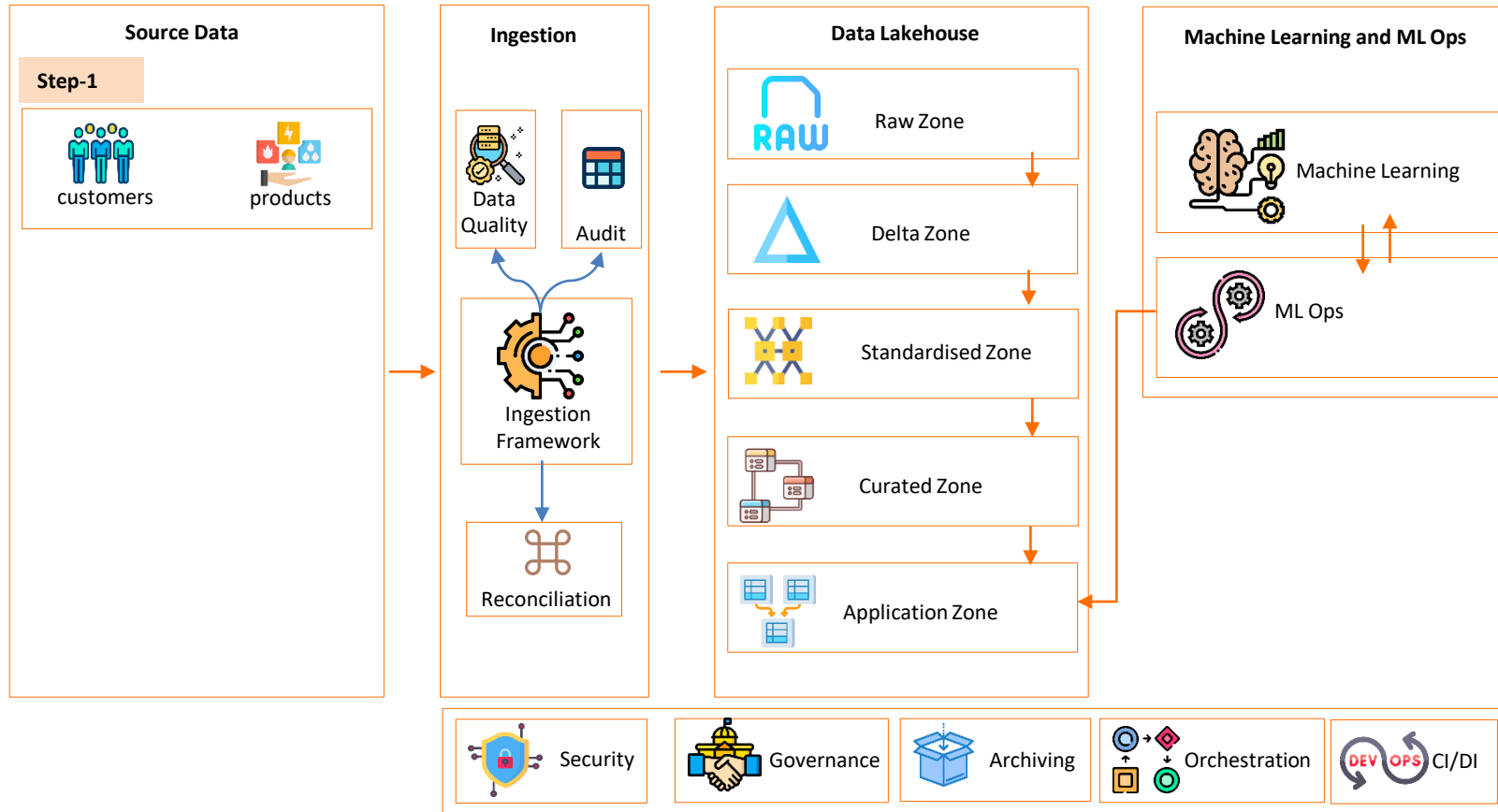


Illustration of Data flow

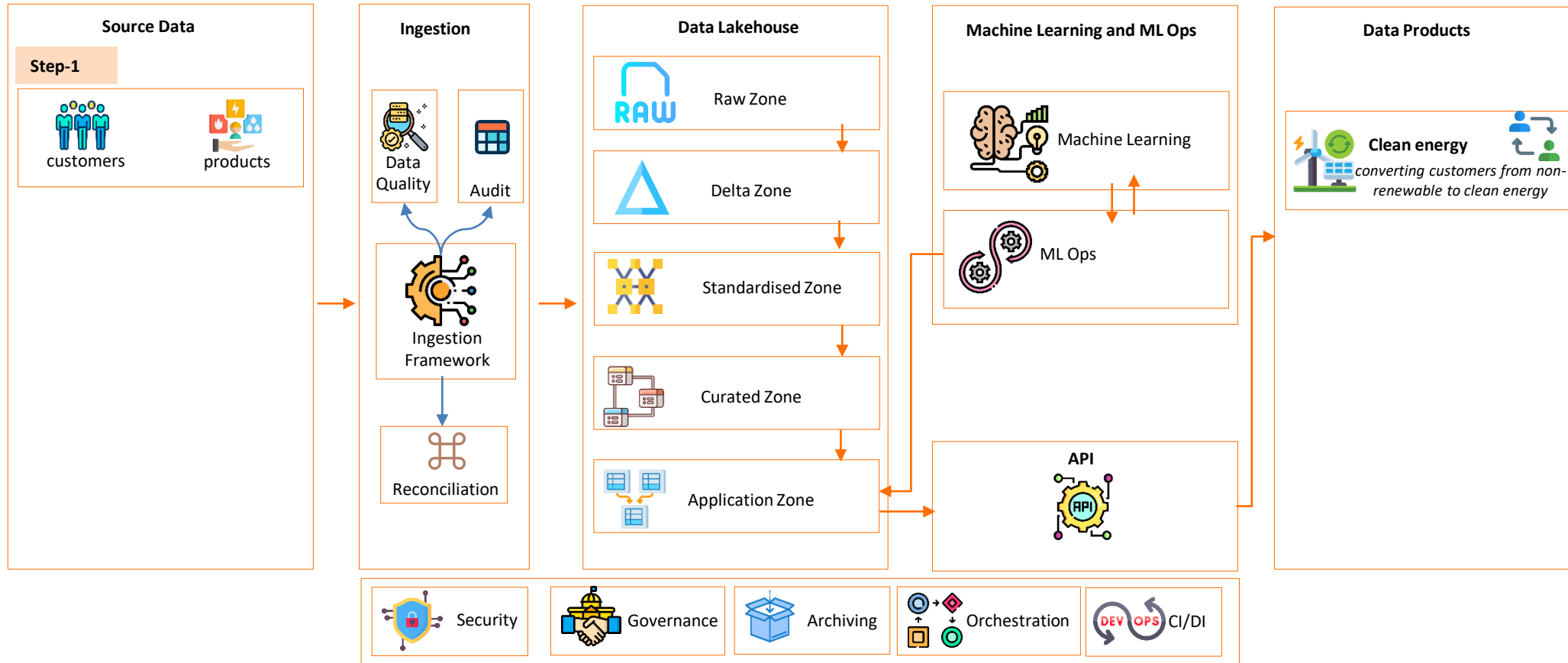


Illustration of Data flow

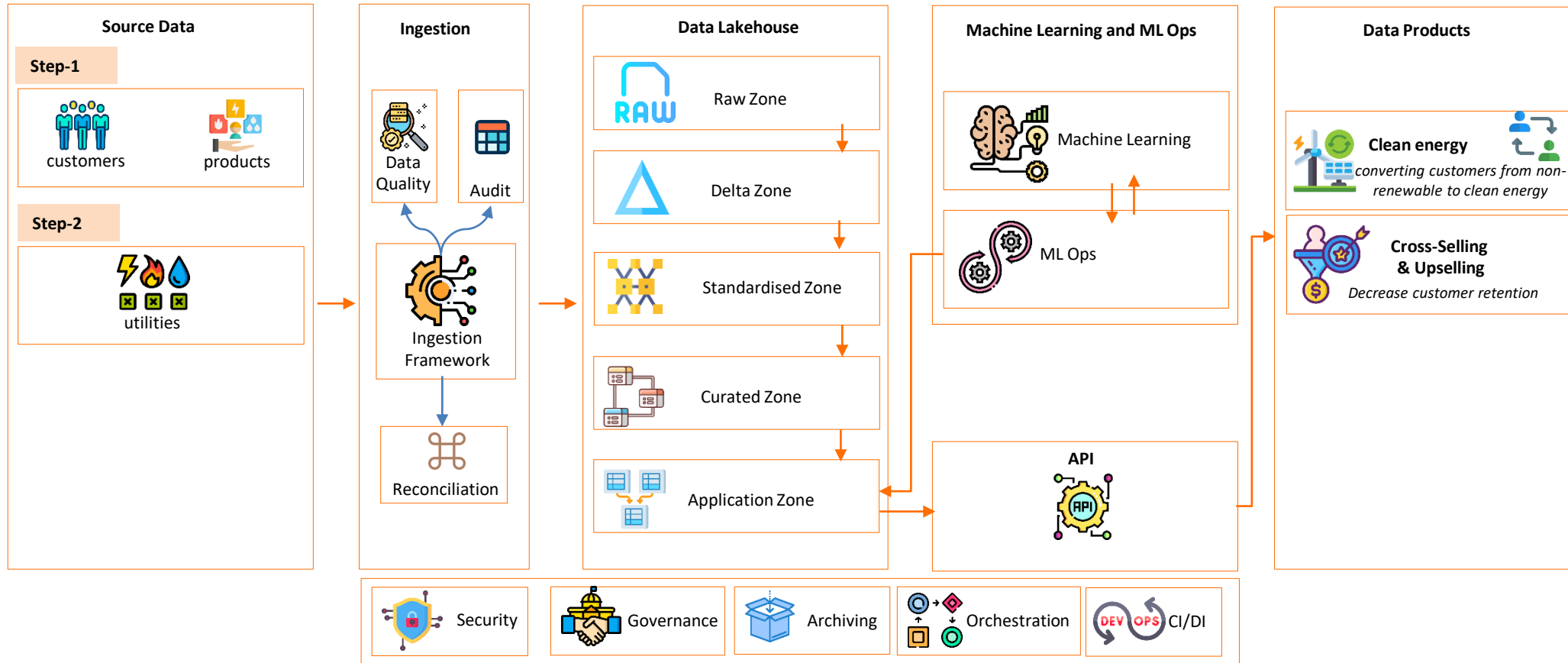
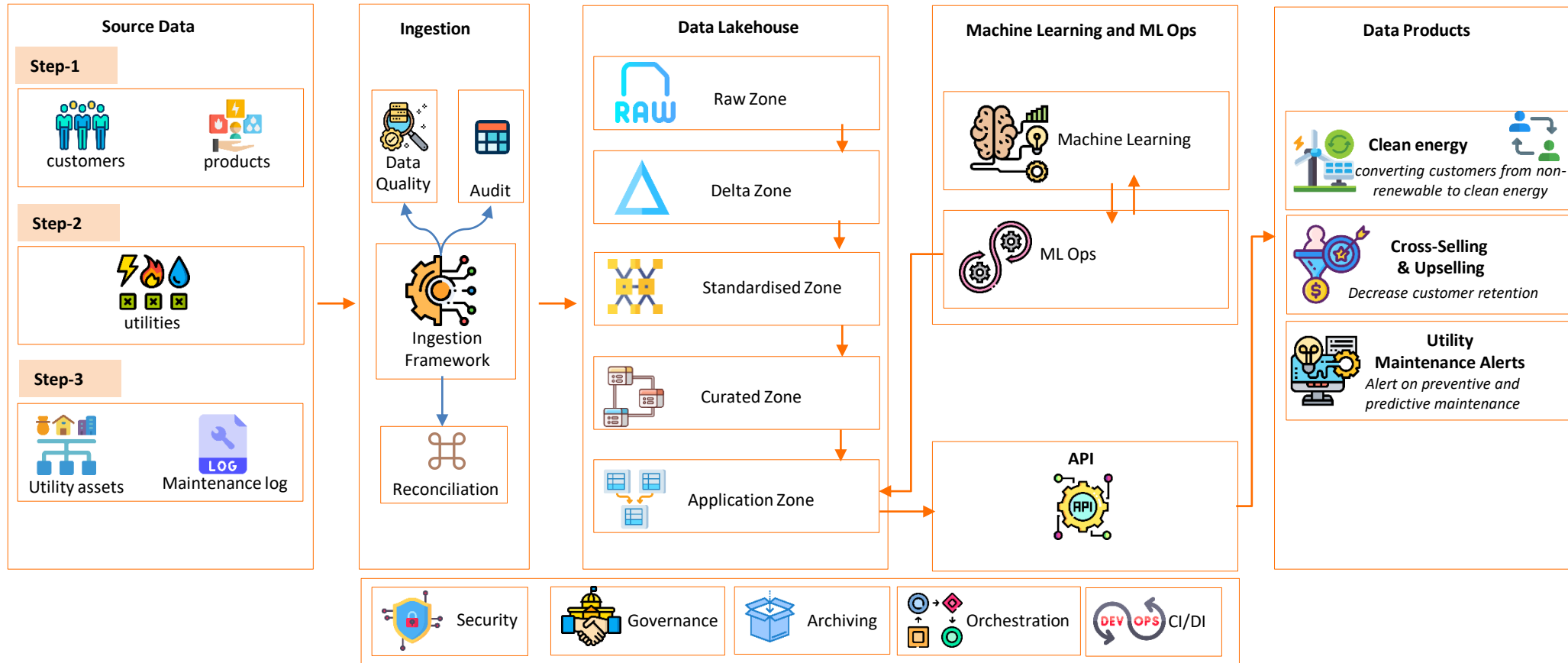


Illustration of Data flow



Data-Driven Digital Transformation with a use-case approach

- ✓ Build and grow data lakehouse – one step at a time
- ✓ Data-driven Ingestion Framework
- ✓ Don't Strive for Perfection
- ✓ Govern with Openness

Understand your data....

build **trust** and **activate** your data....

with purpose to drive **measurable business impact**

Thank you