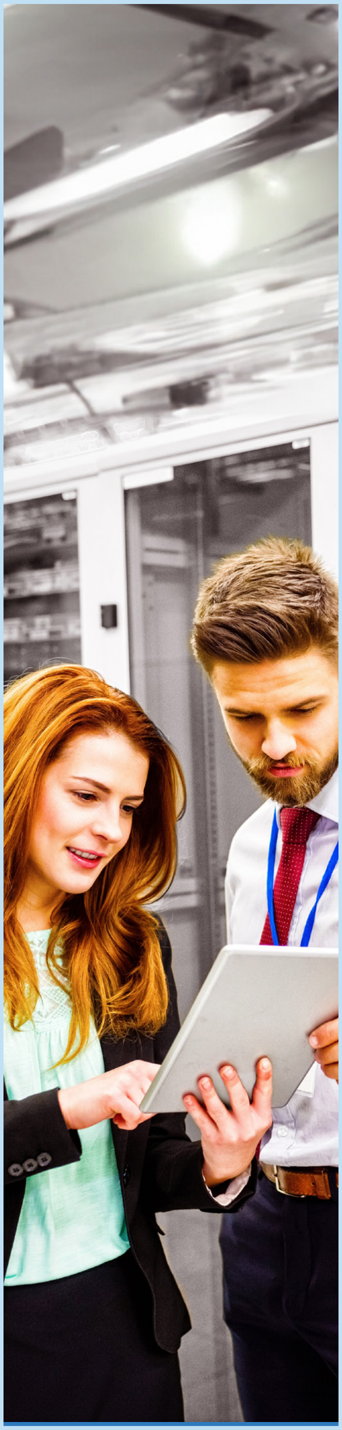
VIEW POINT



DATA LAKE IMPLEMENTATION FOR COMMODITY TRADING AND RISK MANAGEMENT

Abstract

Commodity trading is expected to grow at an exponential rate over the next few years. This will in turn increase the volume of trade, supply chains, and macroeconomic data being generated. Businesses are using several commodity trading and risk management software applications to manage activities such as trading, risk management, and compliance. However, most such packages are not inherently designed to deal with large volumes of raw data.

This paper introduces the concept of a data lake as a consolidated repository for commodity trading data. It discusses data-related issues involved in commodity trading, the challenges faced in implementing a data lake within an enterprise, and suggestions on how to overcome these challenges.

Infosys®
Navigate your next

Introduction to Commodity Trading

Commodity trading is expected to grow at a rate of **2.69%**[1] between 2025 to 2029. Despite global uncertainties, it generated more than US $100 billion[2] in earnings before interest and taxes (EBIT) in 2023

Commodity trading involves trading of agricultural and energy commodities, or metals and minerals in a highly volatile environment. Physical commodities that go into food, electricity, and clothing, and those that enable transportation are an important part of everyday human life.

Trading companies identify the gap in demand and supply of these commodities and try to bridge it with legal contractual agreements. These agreements are in the form of physical contracts, derivative contracts, broker service contracts, or vessel contracts. They capture information such as commodity, locations, currency, quality, grade, pricing, costs, and governing laws along with other details.

Commodity trading and risk management (CTRM) software is used for contract capture, procurement, coordination, pricing, accounting, inventory, scheduling, and risk management. The number of contracts is expected to reach 5,707.00 million[3] by 2029 which will increase the volume of data generated by CTRM systems. 87% of data-driven trading companies in Europe collectively made more than €100 million EBIT[4] in 2022. All these figures point to the importance of data and data-driven decision-making in commodity trading.

Data Related Issues in Commodity Trading

Commodities such as cotton, rice, sugar, wheat, grains, and oilseeds are traded by commodity traders operating in multiple geographies while leveraging different CTRM applications to facilitate trading. Using multiple trading applications can cause the following issues:

- Decentralized data leads to lack of visibility into the consolidated business performance.
- Terminologies with multiple meanings used without context can lead to misinterpretation. For example, the term quantity cannot be analyzed without the context of the report it is used in.
    - Quantity in a trading report refers to contract quantity as represented in Table 1.

Trade Report	
Contract ID	Quantity
A00000	1000

Table 1: Trading report

    - Quantity in a trade operations report refers to the executed quantity. For example, in Table 2, the total executed quantity is 900.

Trade Operations Report	
Contract ID	Quantity
A00000.01	300
A00000.02	400
A00000.03	200

Table 2: Trade operations report

    - Quantity in an open mark-to-market (MTM) report refers to the open quantity that is equal to the difference of contract quantity and the sum of executed quantities.

Open MTM Report	
Contract ID	Quantity
A00000	100

Table 3: Open MTM report

- Differences in data transformation results in data mismatches. For example, using market price with decimal precision of up to two places or up to five places of precision for different reports leads to mismatches during comparisons.
- Lack of governance leads to unauthorized access, posing a security threat to confidential information such as a counterparty's bank account, contact, and address details.
- Decentralized, non-governed, non-normalized and non-organized data affects data quality and impacts business decisions.
- Unorganized data increases storage and processing costs.
- Vast amounts of unorganized data leads to longer data retrieval time.
- Un-harmonized data causes data silos and difficulties in creating relationships across the data. For example, cotton can be represented as Cotton or CTN in different applications leading to reporting confusions.

Data Lake for More Effective Trading

A data lake is a centralized repository of all data in structured or unstructured form stored and accessed using the concept of data catalogs. The main objective of a data lake is to function as a single source of truth from which meaningful, trustable, and real-time insights are generated.

A data lake is ideal for organizations with large volumes of data, multiple operating geographies, disparate applications, and the need for analytics on the data for better decision-making. Qualities of a data lake include stable governance, normalization, security, and effective management of data. This aids data-driven decisions that help boost the bottom line of trading companies.

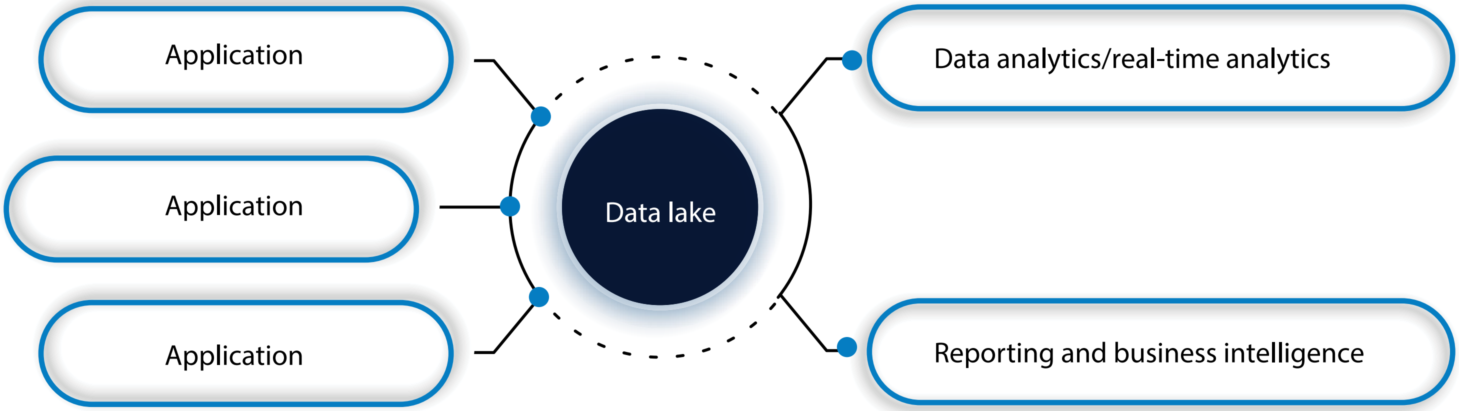

Application
Application
Application
Data lake
Data analytics/real-time analytics
Reporting and business intelligence

Figure 1 – Simplified representation of data lake usage

Benefits of Using a Data Lake

A data lake enables organizations to store huge amounts of structured and unstructured data providing flexibility and scalability for diverse analytics needs. With a data lake, commodity trading firms can:

- Create a centralized data repository for better visibility into the consolidated business performance
- Ensure data governance and security by providing dedicated access controls that prevent unauthorized access to confidential information
- Normalize data for disambiguation of underlying data and help build relationships across systems
- Organize data through classification, cataloging, and metadata management, promoting comparability across applications by providing contextual insights
- Generate impactful business insights owing to the interconnectedness of data across systems
- Achieve enhanced data quality with integrated quality control during the data lake creation process
- Minimize data retrieval time because of the centralization and connectedness of the data lake
- Create uniformity in the data transmission process and compare reports with improved accuracy
- Reduce storage costs by up to 70%[5]

Challenges in Data Lake Implementation

Implementing a data lake in an enterprise is no small project. It involves several challenges spanning technology, leadership buy-in, and implementation quality. However, there are ways to ensure these challenges are overcome and a data lake is created to help the organization scale new business heights.

Executive buy-in

Securing leadership agreement for a data lake project is one of the biggest challenges due to:

- Budget constraints
- Lack of insight into the business benefits of a data lake
- Complexity of implementation

Convincing leadership to invest in a data lake involves:

- Aligning the data lake requirement with the company's vision
- Estimating and projecting the return on investment
- Preparing a phase-wise and incremental plan which provides clear visibility, optimal budget requirements, and project flexibility.
- Projecting the expected overall KPI improvement with research-backed information
- Presenting a benchmarking study of the market leader's or competitors' performance after their data lake implementation

Technology constraints

During the extract, transform, load or extract, load, transform (ETL/ELT) process, there are various challenges involved. These include:

- The diverse nature of data with respect to applications, data structure, data type, and other features
- Decisions on frequency of data refresh, treatment of slowly changing dimensions, and data loading options during ETL (append/update/upsert data)

Some of these technology constraints can be managed by:

- Developing robust ETL processes based on refresh frequency such as EOD, EOM, intraday, and using tools to monitor the health of all ETL processes
- Configuring flags to identify data loss such as implementation of data quality assurance (QA) metrics and data profiling as a part of the ETL process
- Defining data loading options for each dedicated ETL process based on the business requirements. For example, if the requirement is to monitor daily profit and loss, then update loading type is sufficient to view the bottom line

Communication challenges

Data lake implementation requires open communication channels between multiple stakeholders across diverse geographies. Ineffective communication can lead to breakdowns, misinterpretation of requirements, and can affect overall project delivery, quality, and timelines.

Communication challenges can be managed by:

- Implementing standard practices such as project kickoffs and signoffs on requirement and technical specifications
- Preparing meta data management policies for data governance and catalog creation processes for harmonization and data organization
- Conducting joint workshops to arrive at a common understanding and present solutions
- Defining a strong plan with expected deliverables for each phase of the project leading to clear visibility

Quality Challenges

Despite preventive QA checks during the ETL process, issues such as duplicates, data type mismatches, null/blank value generations, data losses, and data errors may be found in the ETL's output which pose a risk to data quality. Handling huge volumes of data is inherently a challenge from a quality perspective.

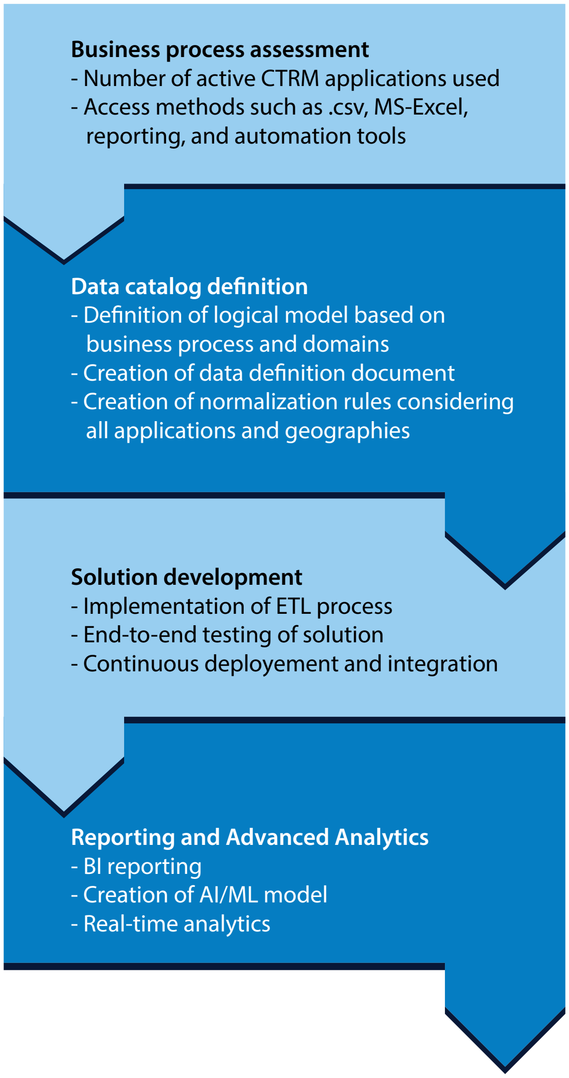
Addressing some of these quality issues involves:

- Reconciling the data lake's output with the source data to identify all quality issues and carrying out a root cause analysis to fix the issues
- Carrying out regression testing to identify failures because of changes in the ETL process
- Standardizing the tools and approach for exploratory data analysis (EDA) and querying and visualization to achieve:
  - » A better understanding of data profiles
  - » Faster identification of root causes of issues

Sample Data Lake Implementation

A sample outline of a data lake implementation strategy for a commodity trading company is shown in Figure 2. The project involves four phases:

- Assessing the as-is state of business processes
- Preparing the data catalog
- Developing the solution
- Designing reporting and advanced analytics tools and dashboards



Business process assessment
- Number of active CTRM applications used
- Access methods such as .csv, MS-Excel, reporting, and automation tools

Data catalog definition
- Definition of logical model based on business process and domains
- Creation of data definition document
- Creation of normalization rules considering all applications and geographies

Solution development
- Implementation of ETL process
- End-to-end testing of solution
- Continuous deployement and integration

Reporting and Advanced Analytics
- BI reporting
- Creation of AI/ML model
- Real-time analytics

Figure 2 – Simplified data lake implementation strategy for a commodity trading company

Data Lake Implementation Case Study

Problem statement

A global agriculture and energy commodity trading company was facing reduced business performance, and data security and quality issues. This was primarily due to their business teams accessing and using their CTRM applications in multiple ungoverned ways.

Infosys solution

Infosys partnered with industry leaders and used best-in-class platforms and technologies to successfully develop and deploy a data lake for the client with features such as data quality, data organization, faster data access and retrieval, and improved data governance. The solution enabled the client to ingest their CTRM data into the data lake to achieve advanced analytics and deep analysis for faster and more empowered decision-making.

Technologies used

Microsoft Azure

Azure SQL

Power Automate

PowerApps

kafka

PostgreSQL

React

.NET Core

Highlights

Some of the notable highlights of the project include:

- Metadata management and governance by centralizing and eliminating direct connection to source applications
- Standardized and faster access to data using automation and by eliminating disparate ways of data access
- Reusability and data comparison between sources by leveraging catalog creation
- Scalability and guaranteed data quality

Conclusion

Commodity trading is on an accelerated growth path and will witness an exponential leap in the volume of data generated in the coming years. While CTRM applications help manage various aspects of commodity trading, these alone will not suffice to handle this immense increase in data.

A data lake acts as a single source of truth into which all data analytics models leveraging artificial intelligence, machine learning, deep learning, and large language models can be built to enable well-informed business decisions.

Using CTRM applications integrated into a data lake is the way forward for commodity trading enterprises to succeed. A data lake empowers commodity trading businesses to integrate and analyze vast, diverse datasets such as market trends, trade records, and supply chain metrics, enabling data-driven decision making and optimized operations.

However, deploying a data lake is fraught with challenges such as technology choices, budget approvals, leadership buy-in, and quality of implementation. Partnering with an expert such as Infosys will help businesses fast-track their data lake solution by leveraging our deep knowledge, industry partnerships, and implementation expertise.

References

1. Commodities - Worldwide | Statista Market Forecast
2. Commodity markets and trading in times of uncertainty | McKinsey
3. Commodities - Worldwide | Statista Market Forecast
4. Commodity markets and trading in times of uncertainty | McKinsey
5. The Benefits of Implementing a Data Lake in Your Business

Authors

Dhanoordaran V

Associate consultant in the Manufacturing Domain Consulting Group with 4 years of overall experience including 1 year in the commodity trading space.

Stephin Samson

Associate consultant in the Manufacturing Domain Consulting Group with 3 years of overall experience including 1 year in the commodity trading space.

Reviewers

Asish Thomas

Principal consultant in the Manufacturing Domain Consulting Group with over 25 years of experience in IT.

Manoj Kumar Gupta

Principal consultant in the Manufacturing Domain Consulting Group with over 19 years of experience in domain consulting.

For more information, contact askus@infosys.com

Infosys®
Navigate your next

Infosys.com | NYSE: INFY

Stay Connected