

AI-Powered Digital Regulatory Reporting

Quality and Compliance Masters
Columbia University + Databricks + JWG-IT + Regnosys

PRESENTERS: Keyi Wang, Charlie Shen, Colin Lin, Jordan Kramer

DATE: Feb 7, 2025

DEPARTMENT: N/a

AGENDA

- ▶ Problem Statement
- ▶ Solution Overview
- ▶ Technical Implementation
- ▶ Architecture
- ▶ Impact and Demo

- ▶ Digital Regulatory Reporting is messy
 - Every agency has its own reporting format
 - In addition to required fields, there are rules that must be enforced on and across columns
- ▶ Fixing it is costly; not fixing it is much worse
 - Once institutions identify errors, there is a manual process for reviewing and correcting them.
 - For errors that aren't corrected, non-compliance can cost billions.
- ▶ CDM is better, but inflexible
 - In cases where trade events are malformed, we might have the data we need, but it's difficult to write ETL code that's flexible enough to find it without human intervention.

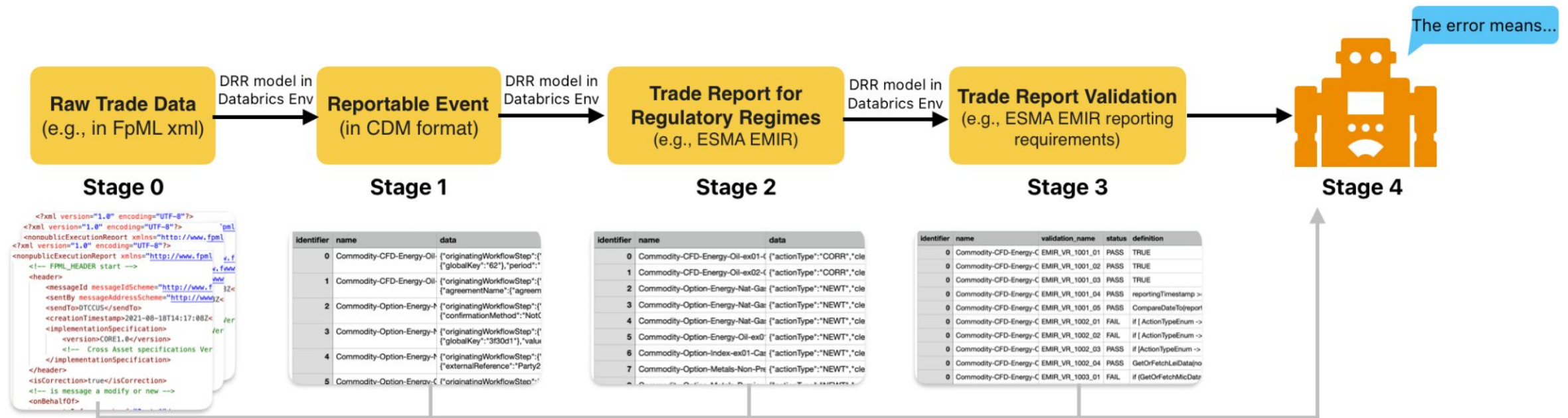
▶ Use the data we have

- Since events can be malformed but still contain the data we need, we can utilize a **chatbot** to add schema flexibility.
- Take advantage of the open source reporting code from FINOS to identify problems with the data, then ask the chatbot to correct it.

▶ Solve simple problems with tech

- There will likely still be errors that the chatbot can't fix, but we can reduce the amount of human intervention required by eliminating simple errors where possible.
- Given that 8-9% of the records returned a failure from our CDM dataset, we expect that our solution can greatly accelerate productivity of compliance specialists tasked with addressing failures.

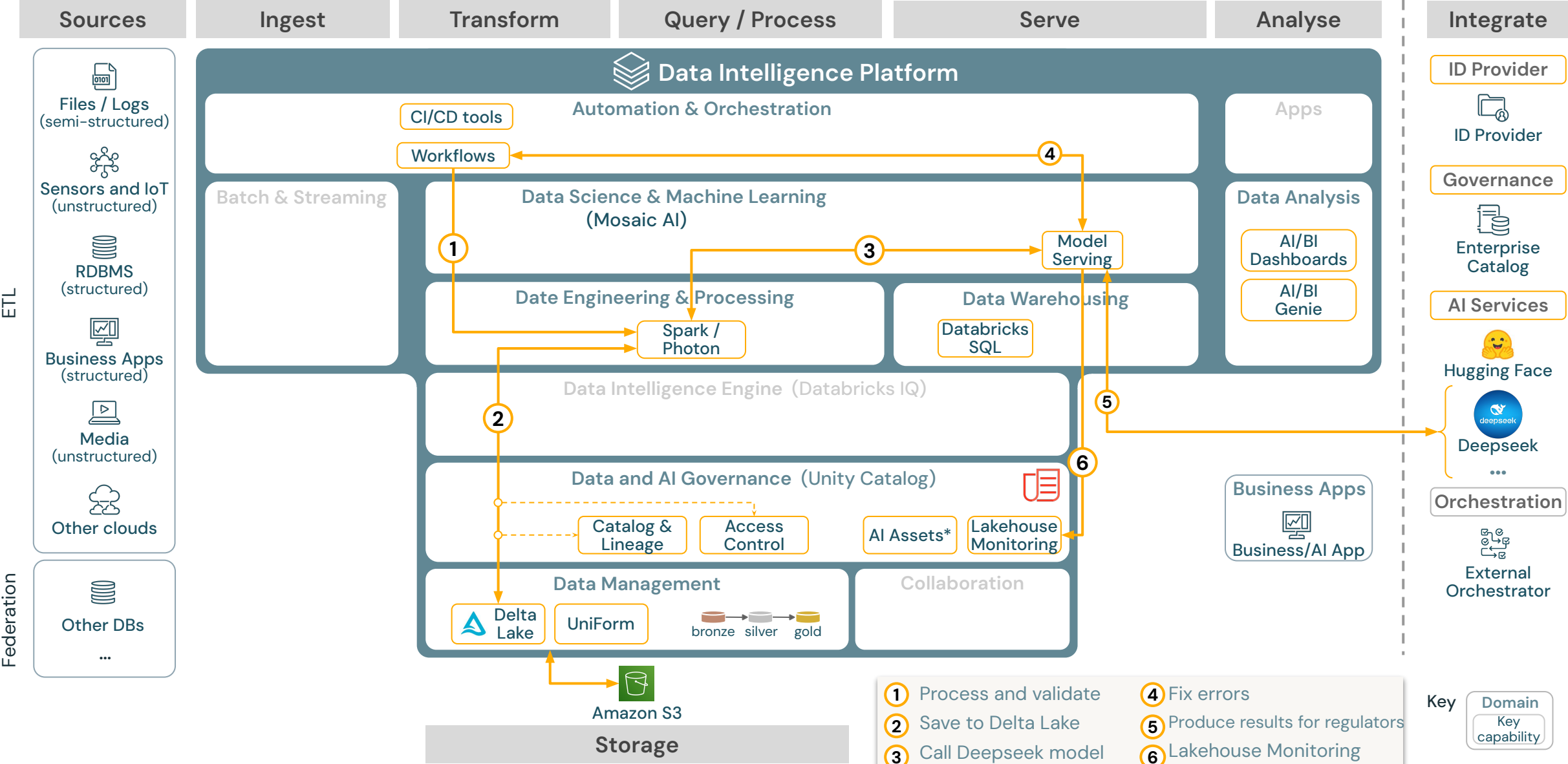
- ▶ **Tech Stack Overview:** DRR model in Databricks, Deepseek R1, Python notebook
- ▶ **Data Usage:** Sample raw trade data and CDM reportable events
- ▶ **Model Performance and Results:** Human evaluation for now, but can validate the corrected version in the Databricks again.





Demo

Technical Implementation



▶ **Potential Path to Production**

- Automated Error Resolution → Implement a feedback loop where fixes update the dataset dynamically.
- Parallel Processing → Scale validation and correction pipelines using distributed computing (e.g., Databricks).

▶ **Scalability** of your solution at an enterprise level

- Automated Error Correction → Implement a feedback loop where LLM-suggested fixes dynamically update datasets, reducing manual intervention.
- High-Throughput Validation → Use distributed computing to process large trade volumes efficiently.

▶ Highlight **collaboration** with industry partners (if any)

- Regulatory Agencies → Align with ESMA, CFTC, and FCA compliance requirements.

▶ Potential **challenges and limitations**




- Deepseek R1 is currently not available for completion on Databricks, only chat.

Conclusion

- ▶ **By using existing open source libraries such as CDM, with cutting edge models like Deepseek R1, and Databricks in conjunction, firms like DTCC can reduce risk, increase efficiency, and meet t+1 goals by addressing regulatory reporting requirements faster**

Appendix

Judging Criteria and Scoring

	Metric	Title	Maximum Score
<div>Technical Innovation</div> 	Innovation and Originality of Solution Design	Does the solution have a creative approach to solve the problem? Was AI used effectively to derive the solution? Judges will evaluate the uniqueness of the solution, the creative approach to solving the use case.	15
	Technical Implementation	The quality and sophistication of the solution's architecture, design, tech stack, industry frameworks and efficiency of the solution.	10
	Solution Viability	Solution's potential for operationalization and enterprise use.	10
	Data Usage and Management	The effective use of provided datasets or external data sources, data preprocessing and feature engineering techniques, and handling of data privacy and security concerns will be evaluated.	10
	Ease of Use	Does it have a user-friendly interface? Can a user use the product with minimum instructions.	5
<div>Value Proposition</div> 	Industry Impact and Viability	Potential value to the Capital Markets industry broadly, market scalability, and revenue generation or cost-saving potential of the solution.	20
	Industry Collaboration <i>(Bonus for "super teams")</i>	Does the implementation (or future phases of the implementation) facilitate/require collaboration across at least 2 or more industry participants? (e.g., through data sharing or other methodologies).	15
<div>Final Presentation and Demo</div> 	Clear Presentation of Problem and Solution	A clear explanation of the problem and solution, the quality of the pitch or demo, and technical design documentation of the project will be important factors.	15
	Effective handling of Q&A		
	Technical Design Documentation		
	Bonus- visual appeal of presentation material		