SPG 2025 JAIPUR

15th Biennial International Conference & Exposition

ROCK TO CLOUD:
GEO-EXPLORATION
EMPOWERING
ENERGY EVOLUTION

SOCIETY OF PETROLEUM GEOPHYSICISTS INDIA

in association with

slb

introducing for the first time

Hackathon

Competition to Collaboration

For More Details :

https://spgindia.org/spg2025/hackathon-2025

Registration Extended Till:

17th September 2025

Open to Professionals, Students and Others

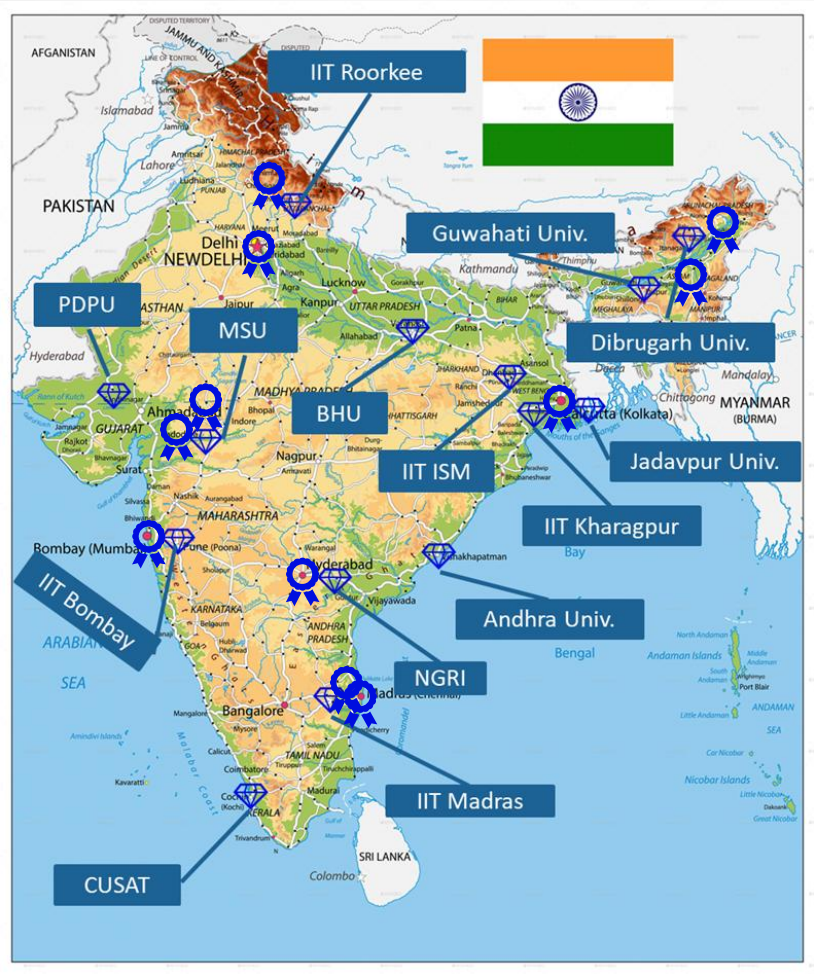
Agenda

- → About SPG
- → About SLB Innovation Factori
- → About Hackathon
- → Introduction to AI/ML Platform
- → Use case
- → LLMs

Society of Petroleum Geophysicists, India
(Regd. No. - 746/ 1992 - 1993)

SPG Today

ROCK TO CLOUD:
GEO-EXPLORATION
EMPOWERING
ENERGY EVOLUTION



- Founded on August 15, 1992 at Dehradun
- Registered on January 4, 1993
- Present Membership: 3200+
- 11 Regional Chapters
- 16 Student Chapters

Endorsing Society

SEG SOCIETY OF EXPLORATION GEOPHYSICISTS

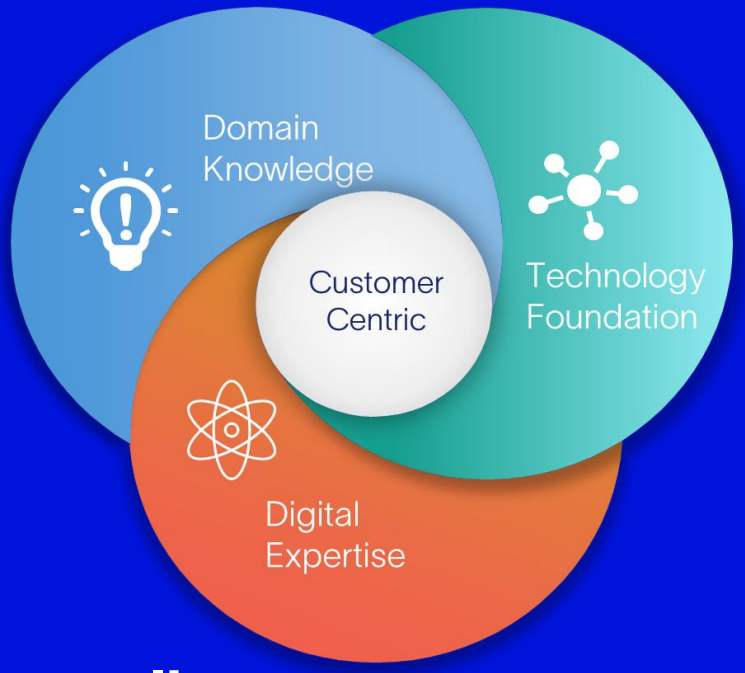
EAGE EUROPEAN ASSOCIATION OF GEOSCIENTISTS & ENGINEERS

Affiliated Society

Australian Society of Exploration Geophysicists

INNOVATION FACTORI

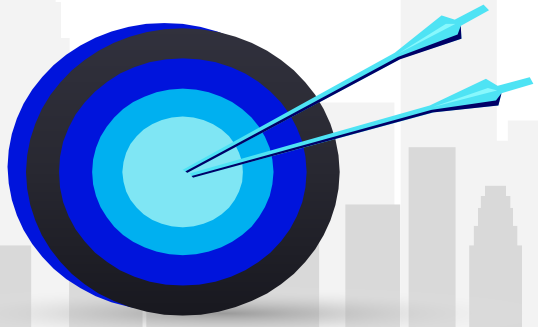
Gateway for collaboration in digital innovation



Global INNOVATION FACTORI network to accelerate and expand collaboration and engagement with the latest in AI and digital solutions

Vision

Mission



Support our customers to **accelerate adoption of AI and innovation** to solve their business challenges

- ✓ Deliver **rapid customer value** in AI and digital customer innovation engagements
- ✓ **Help accelerate customer digital transformation** and business impact of AI and digital solutions
- ✓ Be **close to our customer's digital journey** – be present where they are

Dual-Phase Hybrid Hackathon: From Competition to Collaboration

Innovative hybrid hackathon that brings together students, professionals, data scientists, and geoscientists to tackle AI-driven challenges in subsurface data analysis. This unique two-phase approach transforms traditional competition into collaborative innovation.

Hackathon Overview: A Groundbreaking Format

Phase 1: Virtual Challenge

Teams will work independently on a common dataset, developing their own unique approaches to extract insights from subsurface data.

Phase 2: Collaborative Finals

Top performers are reorganized into diverse, shuffled teams for an intensive collaborative challenge with a fresh dataset.

Who Can Participate?

Open to students, professionals, data scientists, and geoscientists interested in applying AI to subsurface datasets. No prior collaboration between these groups required.

Phase 1: The Challenge

Seismic or well logs interpretation or Insight Generation Using AI

Phase 1: Virtual Screening

Define Problem Statement

Clearly define a specific problem using the provided dataset.

Develop AI Prototype

Build a functional AI prototype to address the problem.

Tool Agnostic

Utilize any tools, including coding frameworks and AI assistants.

Leverage Dataiku

Integrating Dataiku earns extra evaluation marks.

Scope Limitation

Avoid solutions focusing on seismic processing.

Possible Areas of Focus

- Seismic interpretation
- Well log interpretation & processing
- Generating insights from technical reports
- Integrating seismic, well, and report data

Phase 1: Key Expectations & Judging Criteria

Key Expectations

1. Demonstrate innovation, creativity, novelty in your approach.
2. Show real-world impact (e.g., cost reduction, time saving, efficiency, or accuracy improvement).
3. Deliver a working end-to-end prototype of your solution.
4. Present your solution with a clear and concise pitch deck.
5. Include complete code and comprehensive documentation with your submission.

Judging Criteria

1. **Innovation:** How unique and forward-thinking is your solution?
2. **Creativity & Novelty:** Originality in problem-solving and implementation.
3. **Real-World Impact:** Potential for tangible benefits in the industry
4. **Clarity of Pitch Deck:** Effectiveness in communicating your project's value.
5. **Completeness of Solution:** Functionality and robustness of the working prototype.
6. **Use of Dataiku Platform:** Strategic and effective integration of Dataiku features.

Phase 1 Submission Requirements

For Phase 1, each team must submit their solution alongside a comprehensive pitch deck (Max 10 Slides) and adhere to specific expectations designed to foster innovation and real-world impact.

Problem Statement

Clearly define the challenge your team is addressing.

Why It Matters (Need for Solution)

Explain the business or technical importance of your solution.

Proposed Solution

Detail your AI/GenAI-driven approach and methodology.

Data & Tools Used

Provide dataset references (with citations) and list all tools/platforms leveraged.

Prototype Demo/Architecture

Illustrate your workflow, model design, or include screenshots of your prototype.

Value Addition / Impact

Demonstrate how your solution reduces cost, saves time, improves accuracy/efficiency, or adds industry value.

Innovation & Creativity

Highlight what makes your approach unique or novel.

Conclusion & Next Steps

Summarize your results and suggest future improvements or applications.

Phase 1: Submission Details

Pitch Deck

Submit your comprehensive pitch deck, detailing your problem, solution, and impact. Acceptable formats include PDF or PPT.

Code & Documentation

Provide your complete Dataiku project along with thorough documentation explaining your methodology, setup, and results.

Submission Deadline

3rd October 2025

Please ensure all deliverables are submitted by the specified deadline to be eligible for evaluation in Phase 1. Further instructions on the submission portal will be provided closer to the date.

Hackathon Platform: Dataiku

(Supported by SLB)

AI for Everyone

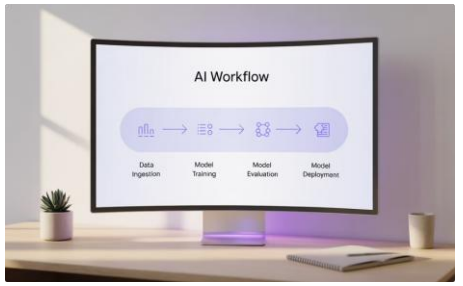
A powerful low-code/no-code platform for building and deploying machine learning models with an intuitive visual interface, integrating Generative AI for rapid iteration on AI solutions.

Accessible to All

Beginner-friendly yet expert-capable, Dataiku supports both technical and non-technical participants, fostering effective collaboration among data scientists, geoscientists, and students.

Comprehensive Training & Support

SLB will provide extensive training sessions and ongoing support covering Dataiku fundamentals, advanced features, and best practices to maximize project potential.



Integrated Environment

Dataiku offers a holistic environment for the entire AI lifecycle, from data ingestion to model deployment, streamlining workflows and fostering rapid experimentation in a fast-paced hackathon.

Evaluation Advantage

Effective and creative use of Dataiku is a key judging criterion, showcasing participants' ability to translate ideas into actionable AI solutions through innovative problem-solving.

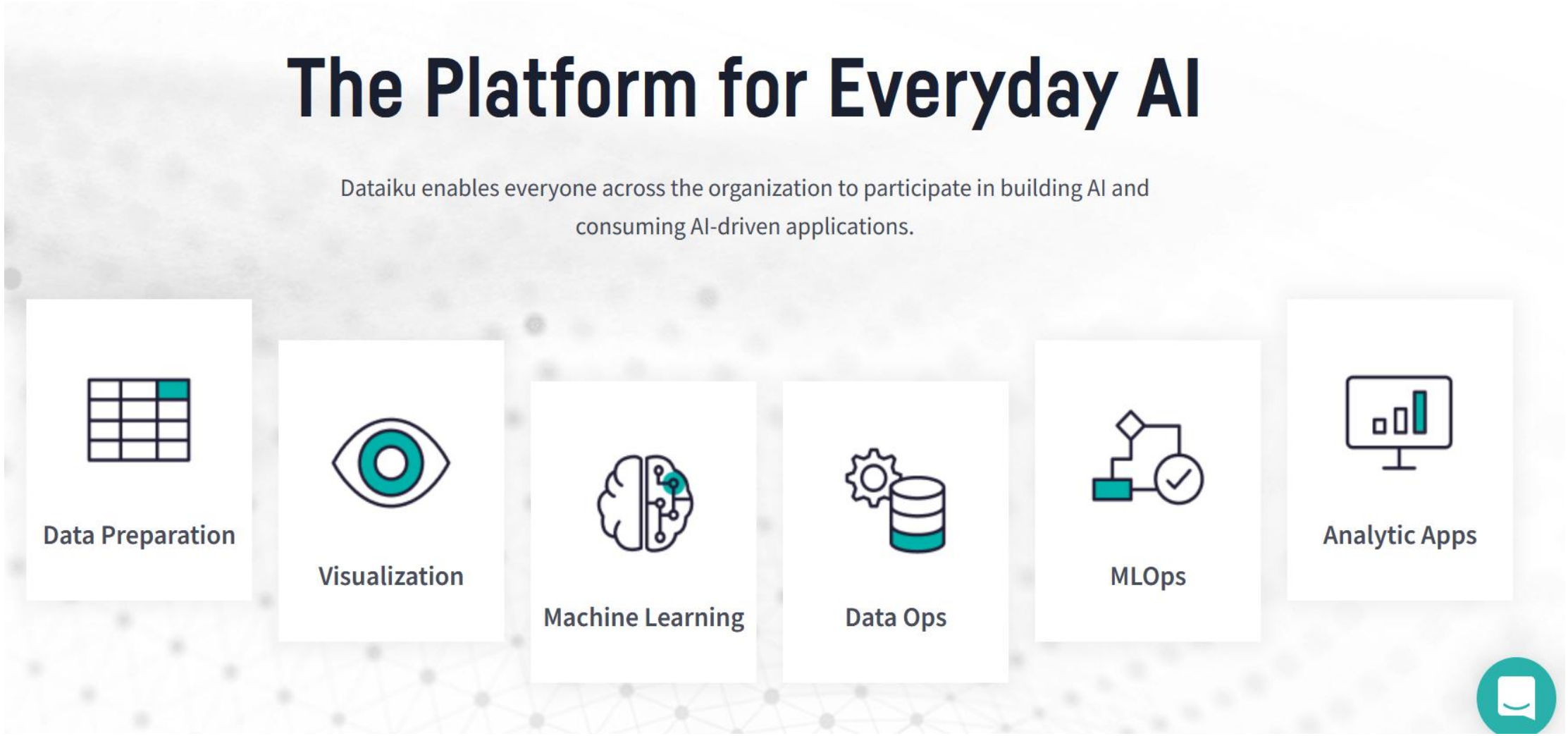
AI/ML platform

An introduction to Dataiku

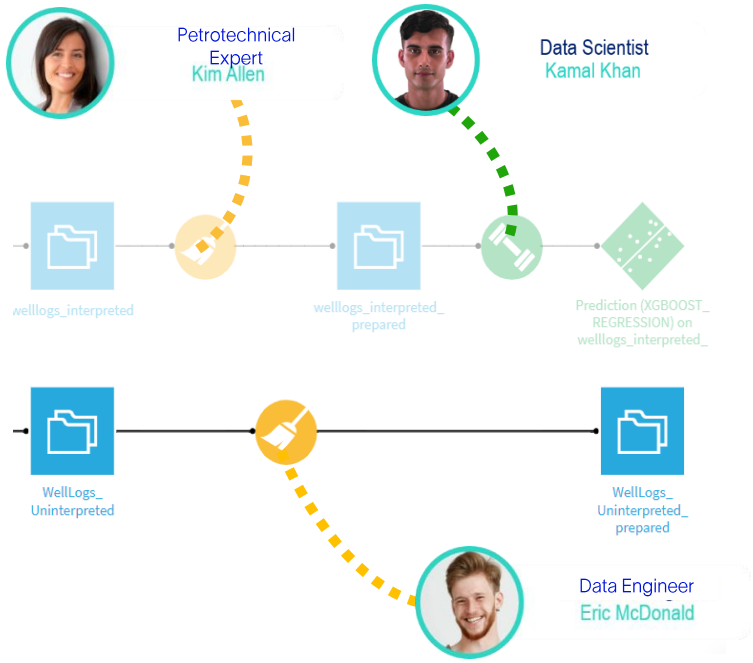
1. Dataiku Presentation
2. DSS concepts
3. Hands-on – Facies Classification

Dataiku presentation

Data Science and Analytics

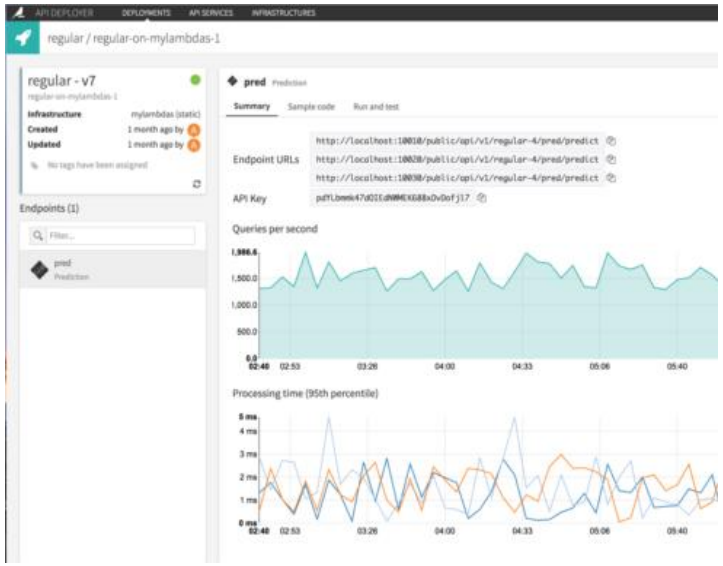


Data Science and Analytics



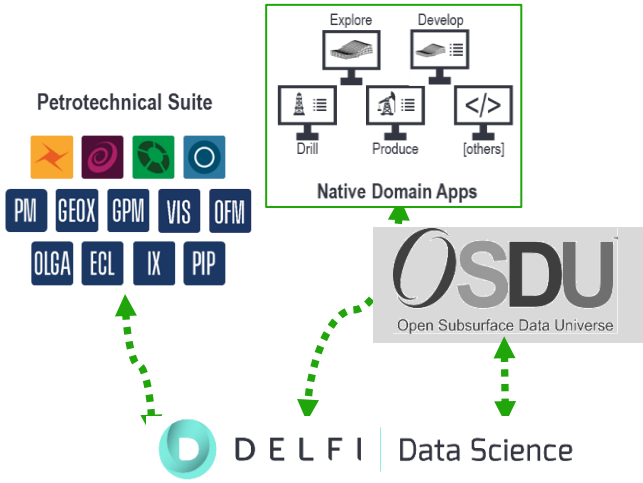
Democratized AI/BI

- AI and BI for everyone
- No code, low code, high code
- Easy to use visual modeling interface



Deployment at Scale

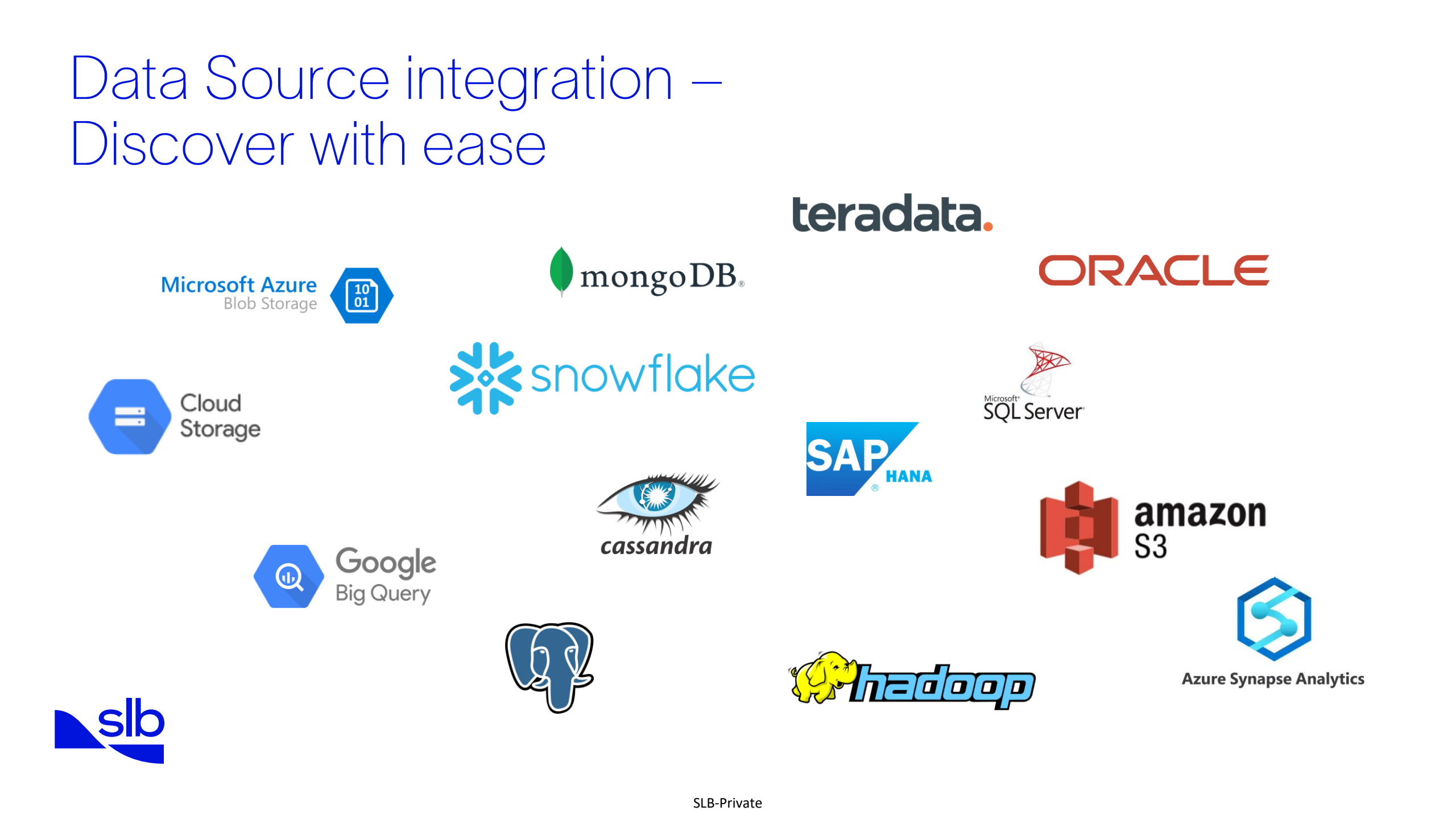
- AI everywhere
- Ease of enterprise deployment
- Multi-cloud deployment



Enrich Workflows

- Integrate to any data source
- Openness to provide integration with applications

Data Source integration – Discover with ease

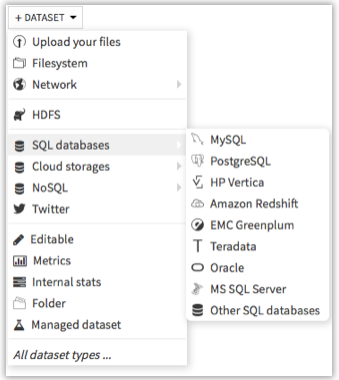


Dataiku DSS

Data Access and Processing

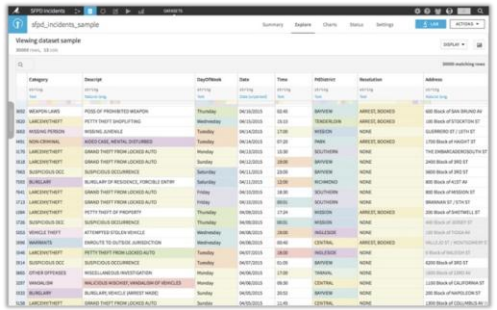
CONNECT TO YOUR (MANY) DATA SOURCES

- **Click based** connection to your datalake, databases, flat files or any other source
- Native connectors for most common technologies (SQL, Hadoop, Spark, NoSQL…)
- Find relevant data with the catalog



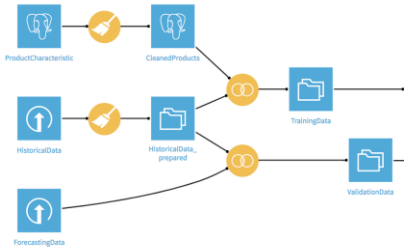
DATA EXPLORATION

- Explore your data easily, investigating potential quality issues.
- Get rich insights into every column with built-in charts and summary information



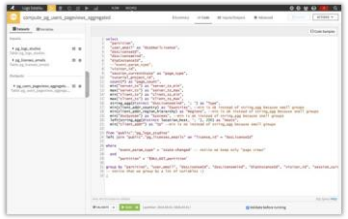
DEVELOP REUSABLE/MAINTAINABLE DATA FLOWS

- Simple representation of overall data processing despite complexity of underlying operations
- Portability of in-built recipes to various execution engines



DATA PREPARATION

- Use the 100+ in-built processors to perform advanced operations in a few click
- Extend the built-in tools by writing custom formulas or code for bespoke transformations.

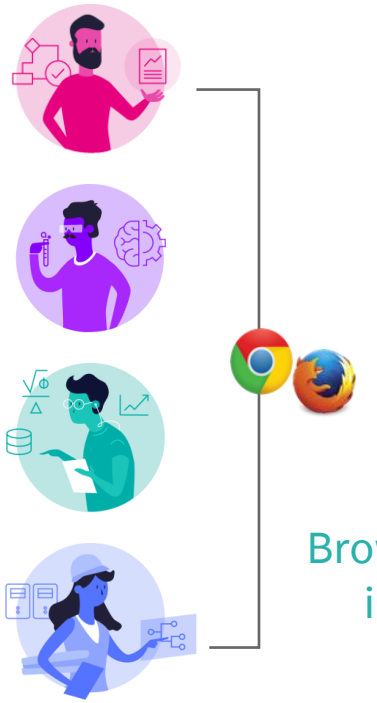


COMMUNICATING INSIGHTS

- Convey insights with 25+ standard charts and map types.
- Easy to integrate with popular visualization solutions like Tableau, Power BI and Qlik

Dataiku DSS

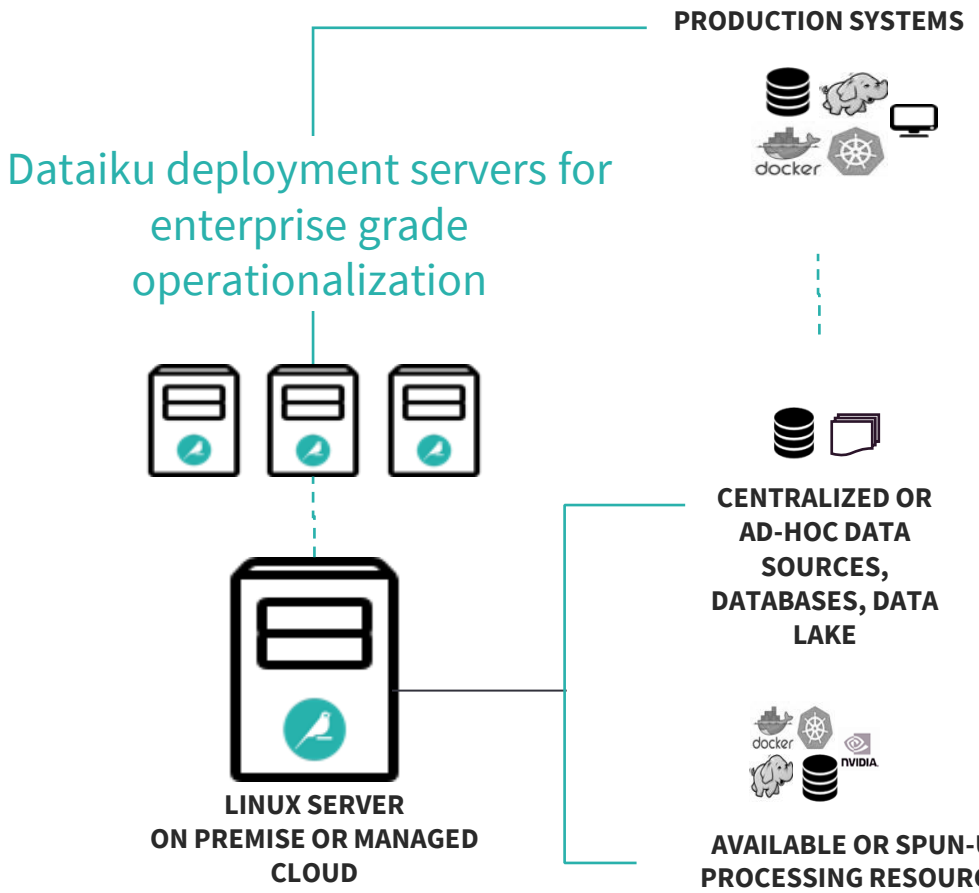
Solution Overview: Architecture



Browser based interface

Dataiku DSS

Solution Overview: Architecture

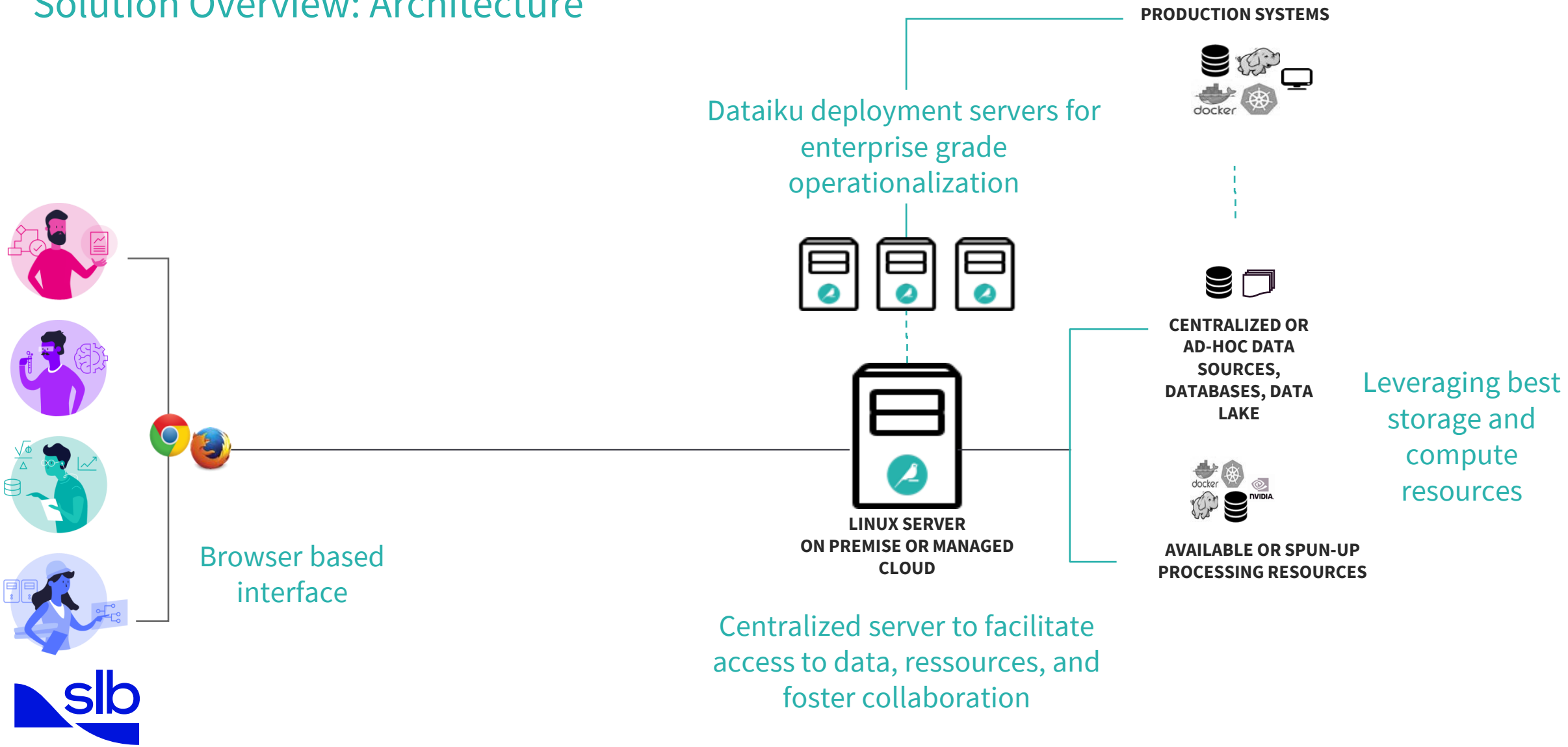


Leveraging best storage and compute resources

Centralized server to facilitate access to data, ressources, and foster collaboration

Dataiku DSS

Solution Overview: Architecture



Dataiku DSS

Solution Overview: Architecture

Browser based interface

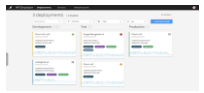
User/task specific interaction modes

VISUAL DEVELOPMENT

VISUALIZATION

COMPLETE CODING ENVIRONMENTS

COLLABORATION AND PROJECT MANAGEMENT



AUDIT, MONITORING AND SCHEDULING

Dataiku deployment servers for enterprise grade operationalization

LINUX SERVER ON PREMISE OR MANAGED CLOUD

Centralized server to facilitate access to data, ressources, and foster collaboration

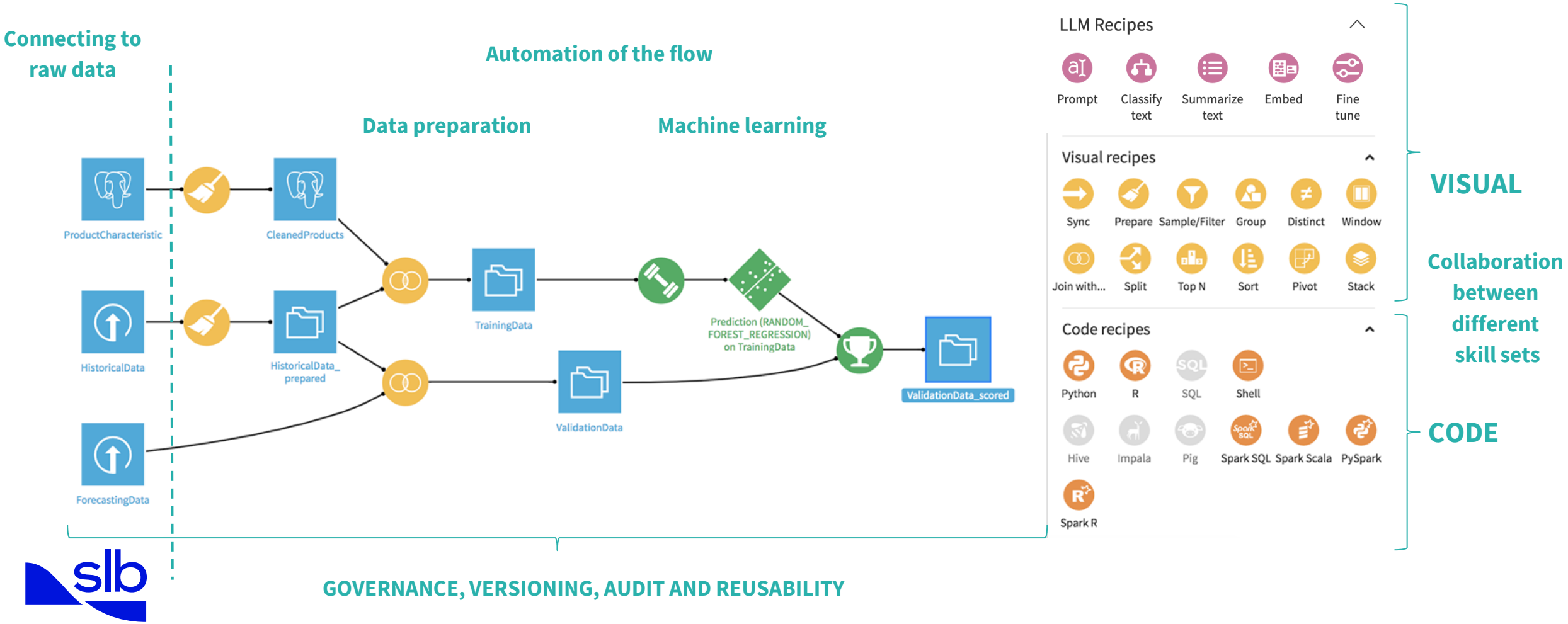
PRODUCTION SYSTEMS

CENTRALIZED OR AD-HOC DATA SOURCES, DATABASES, DATA LAKE

AVAILABLE OR SPUN-UP PROCESSING RESOURCES

Leveraging best storage and compute resources

Inclusive end to end platform



Collaborative platform for data science

Business Analyst

Visual Data Preparation

Auto ML

Data Scientist

R, Python, Notebooks, and many data Science Libraries

Model Monitoring

Analytics Leader

Visualization and validation

Data Engineer

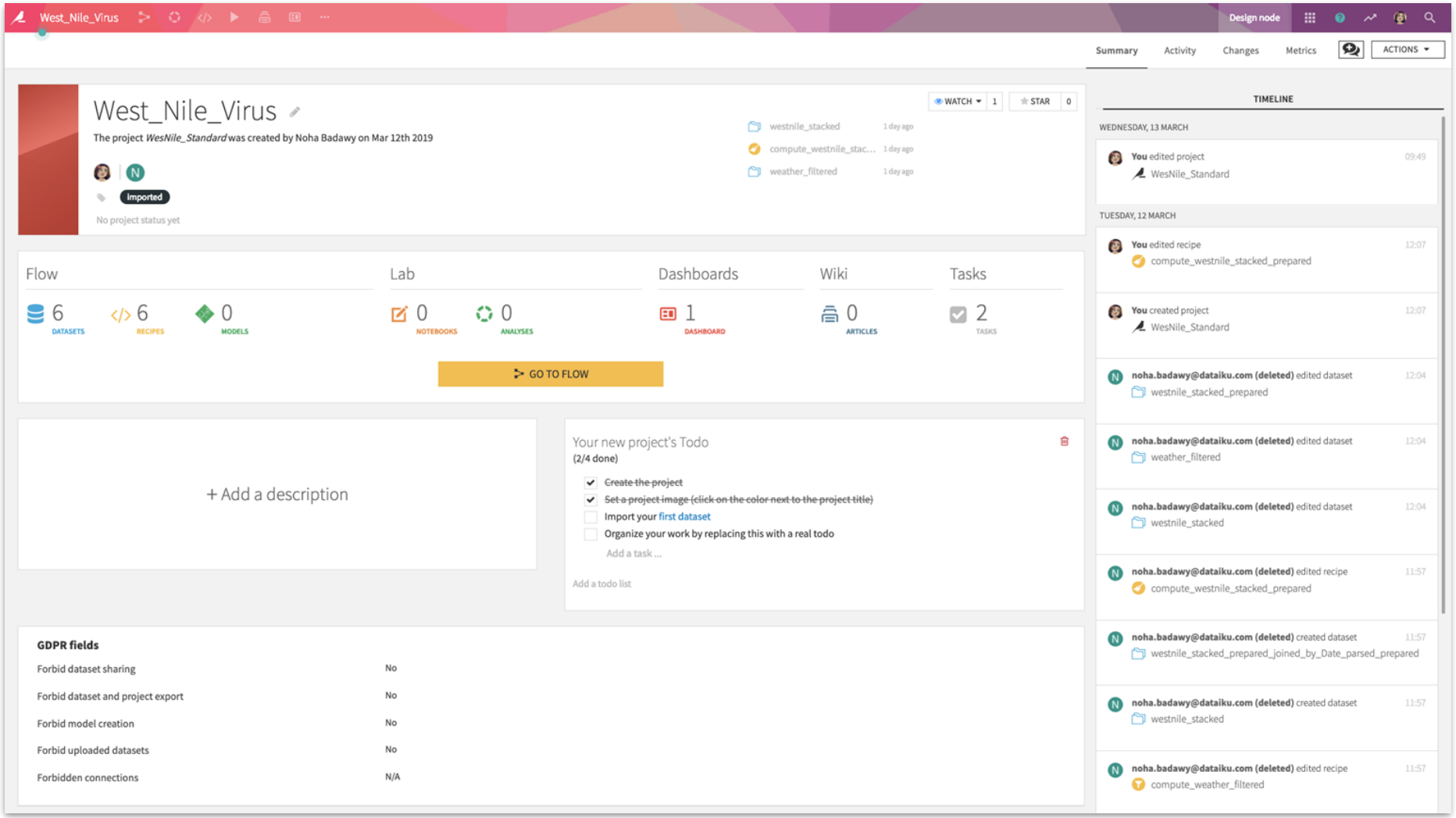
Production deployment

IT Monitoring

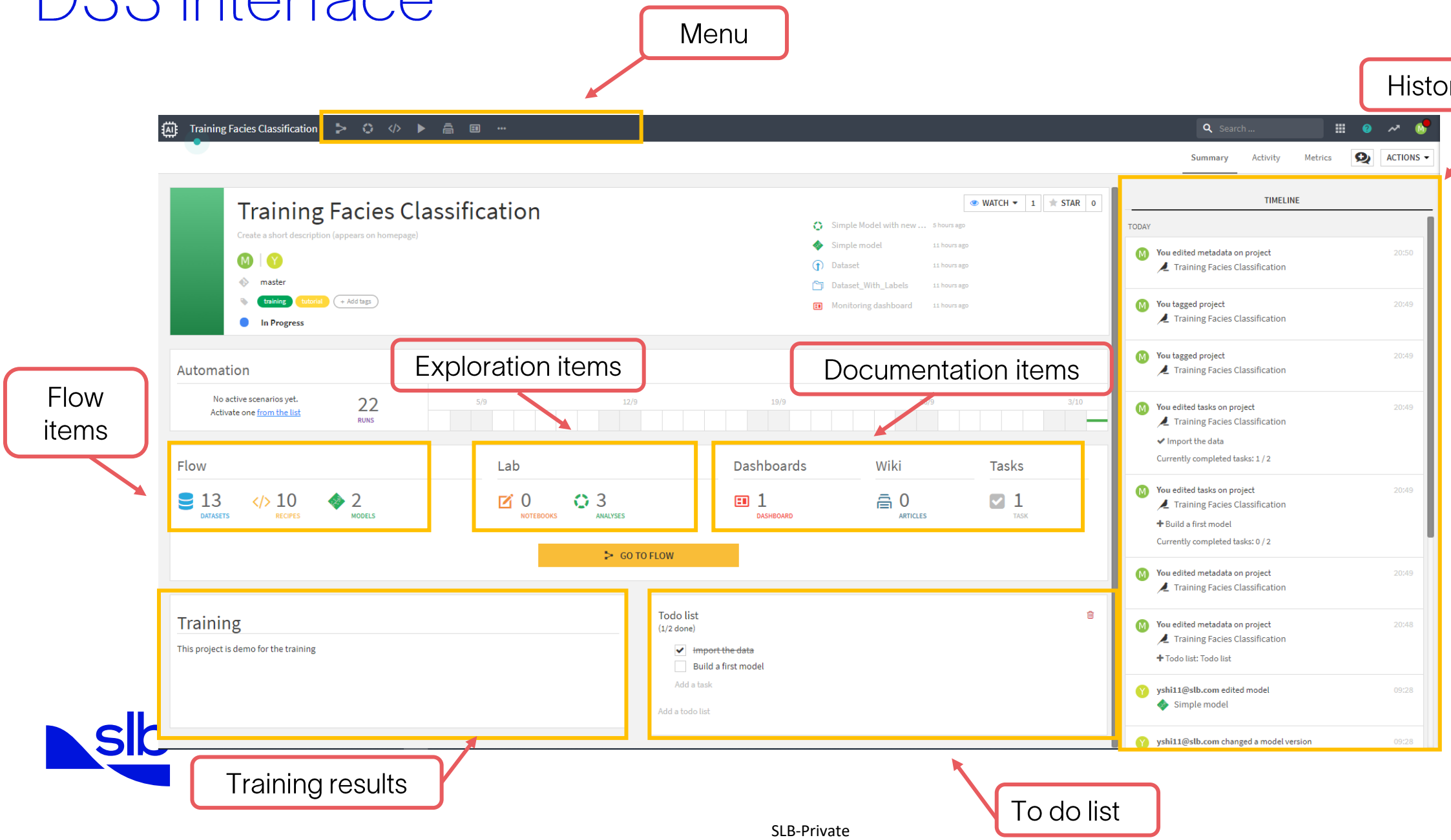
DSS concepts

DSS Concepts

Project homepage

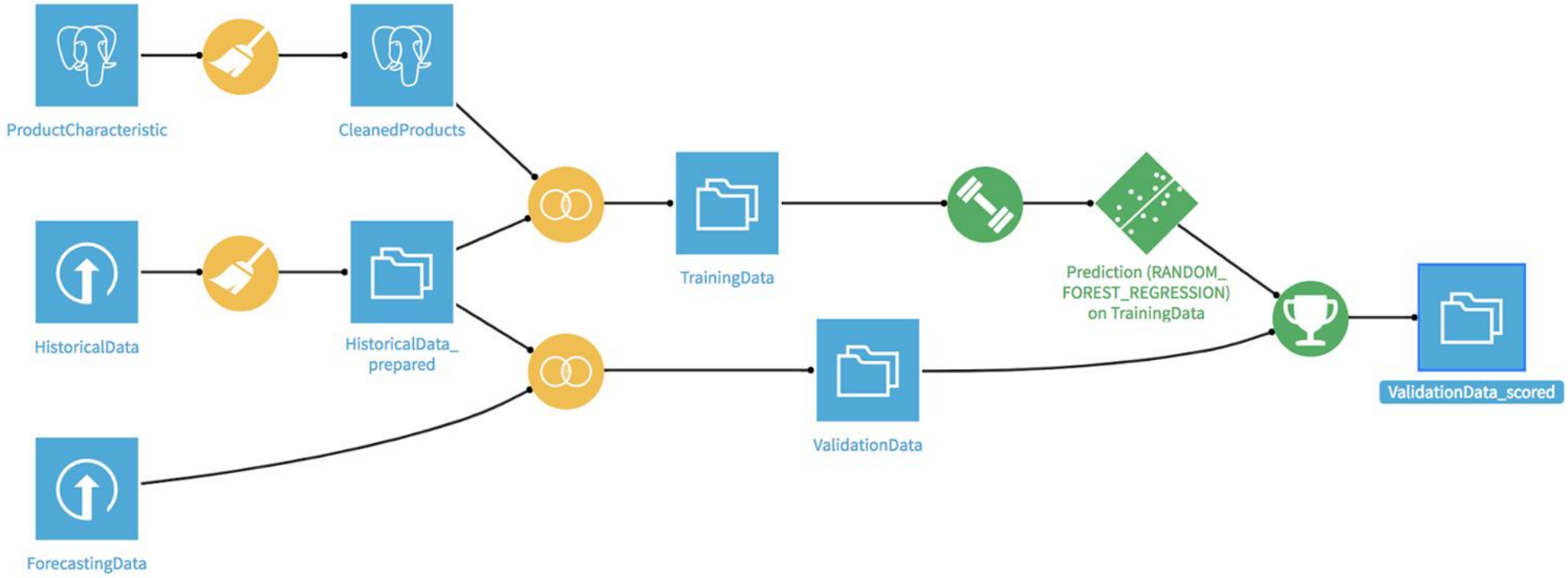


DSS interface



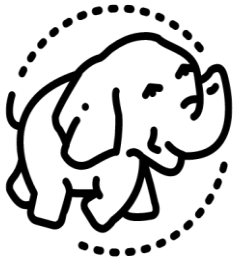
DSS Concepts

Flow, Datasets, Recipes



DSS Concepts

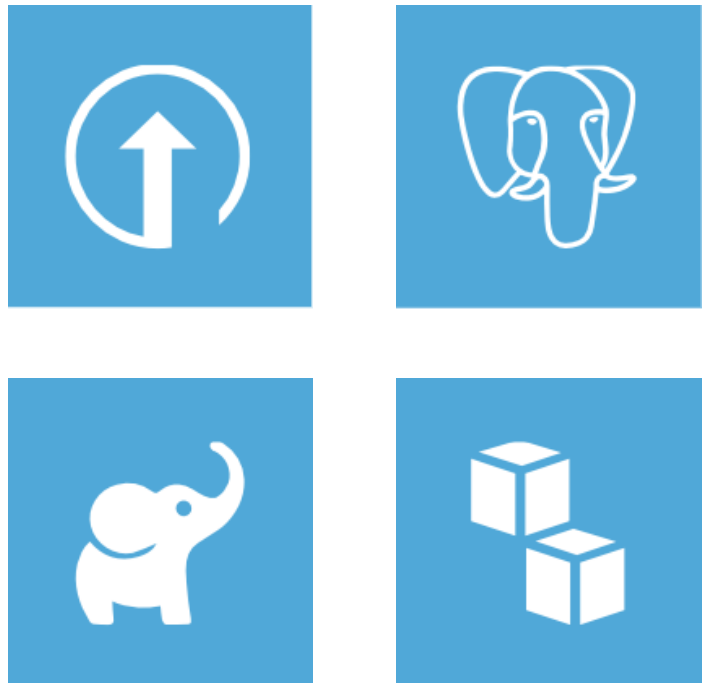
Datasets

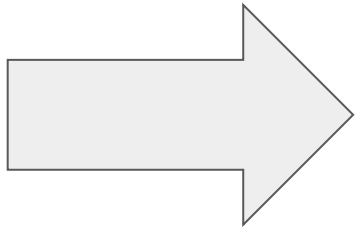


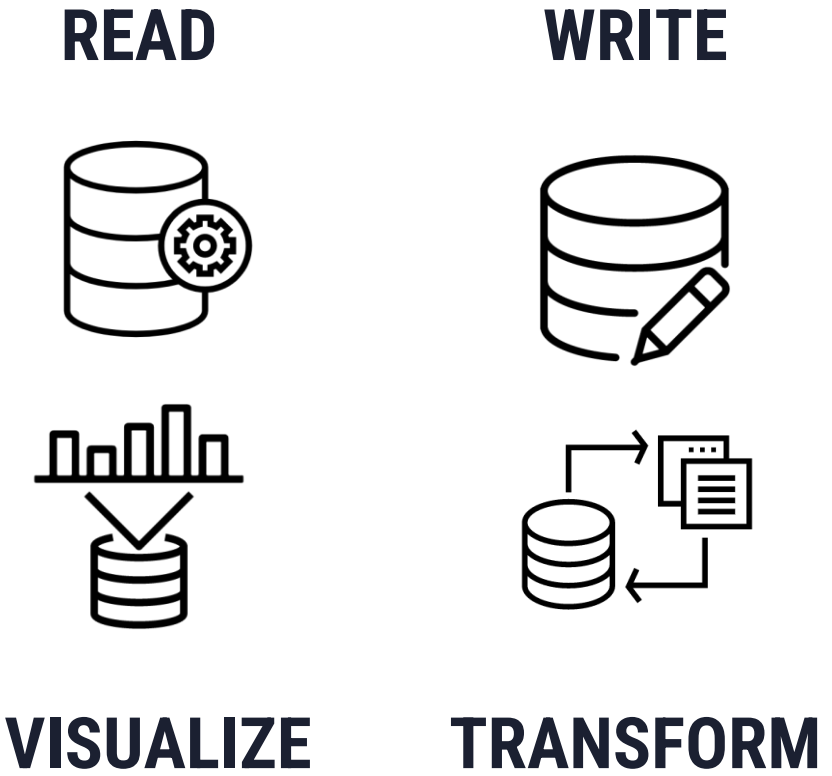
DSS Concepts

Datasets

You can run all the usual data manipulation operations on your datasets, **independently of their connection type**.

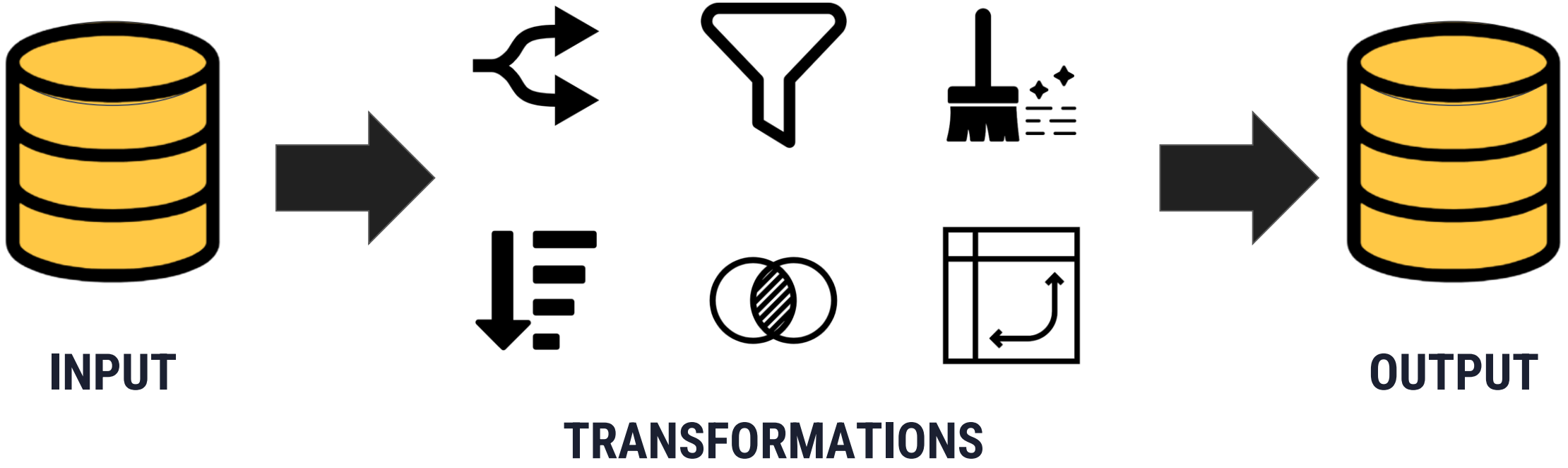






DSS Concepts

Recipes

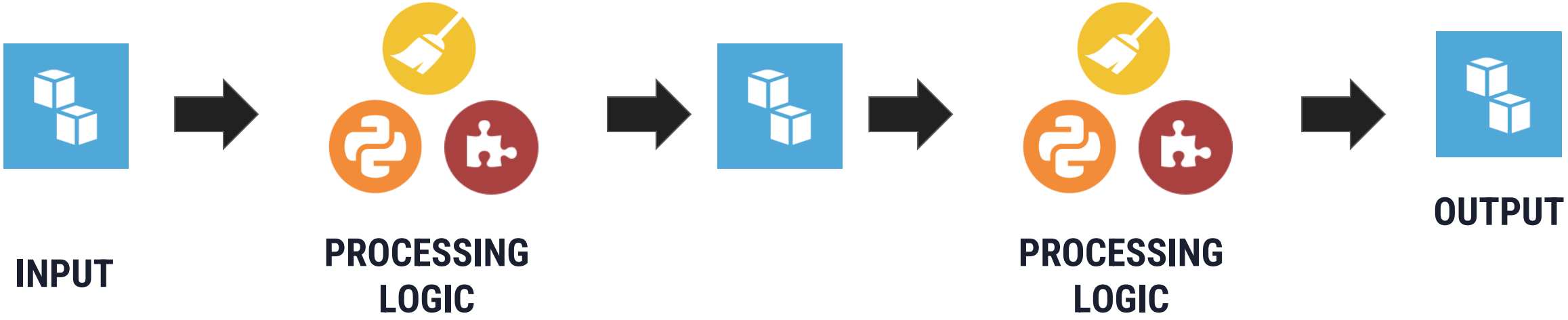


DSS Concepts

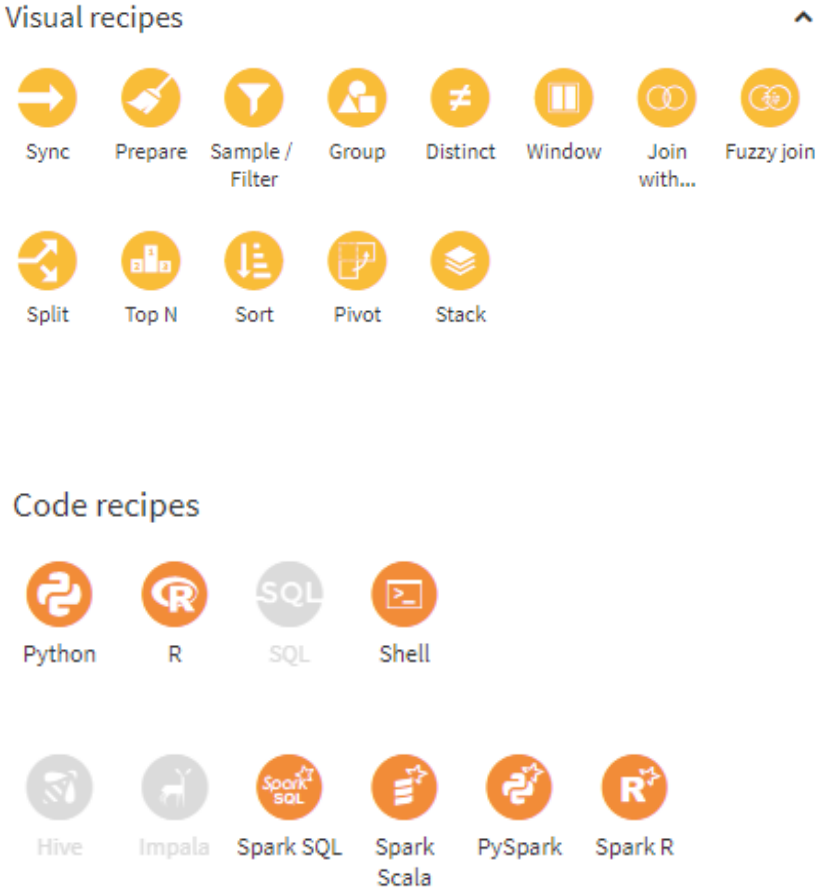
Recipes

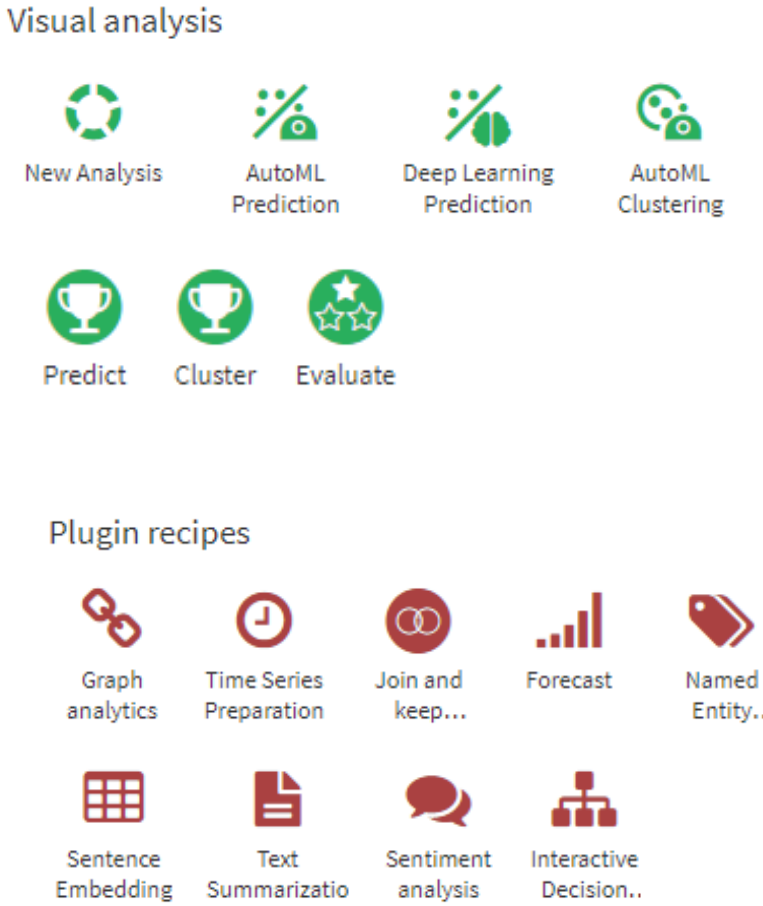
DSS Concepts

The Flow



The recipes





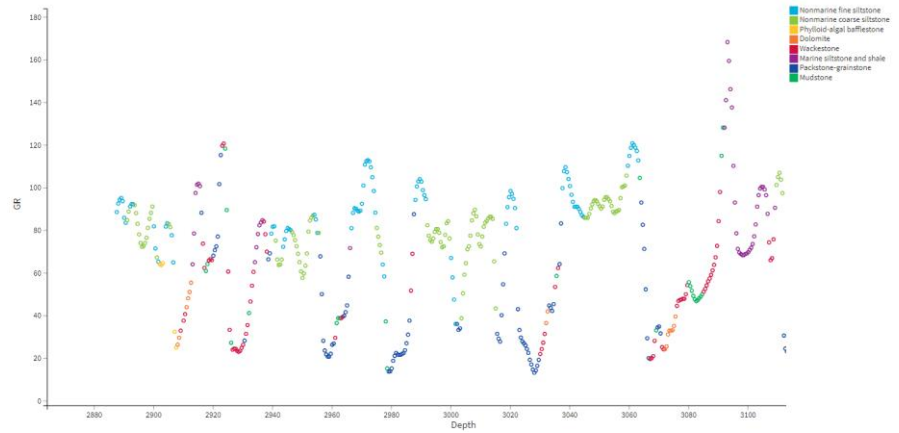
Data preparation

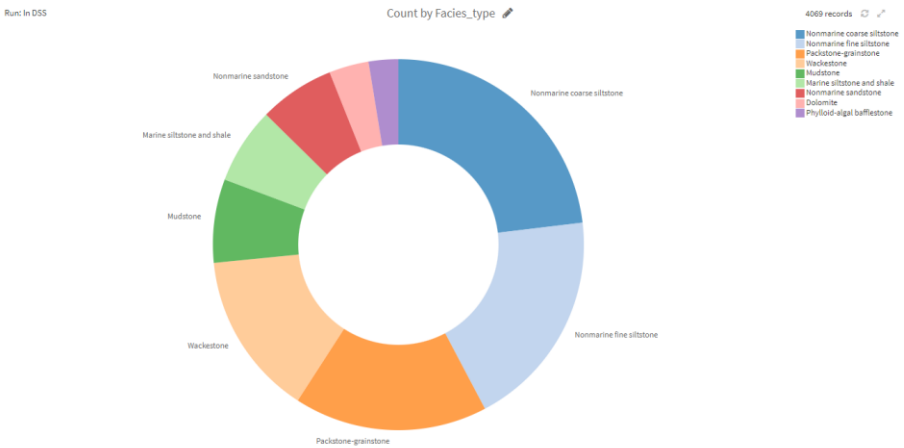
- Dataiku offers many visual tools to prepare, clean, format your data

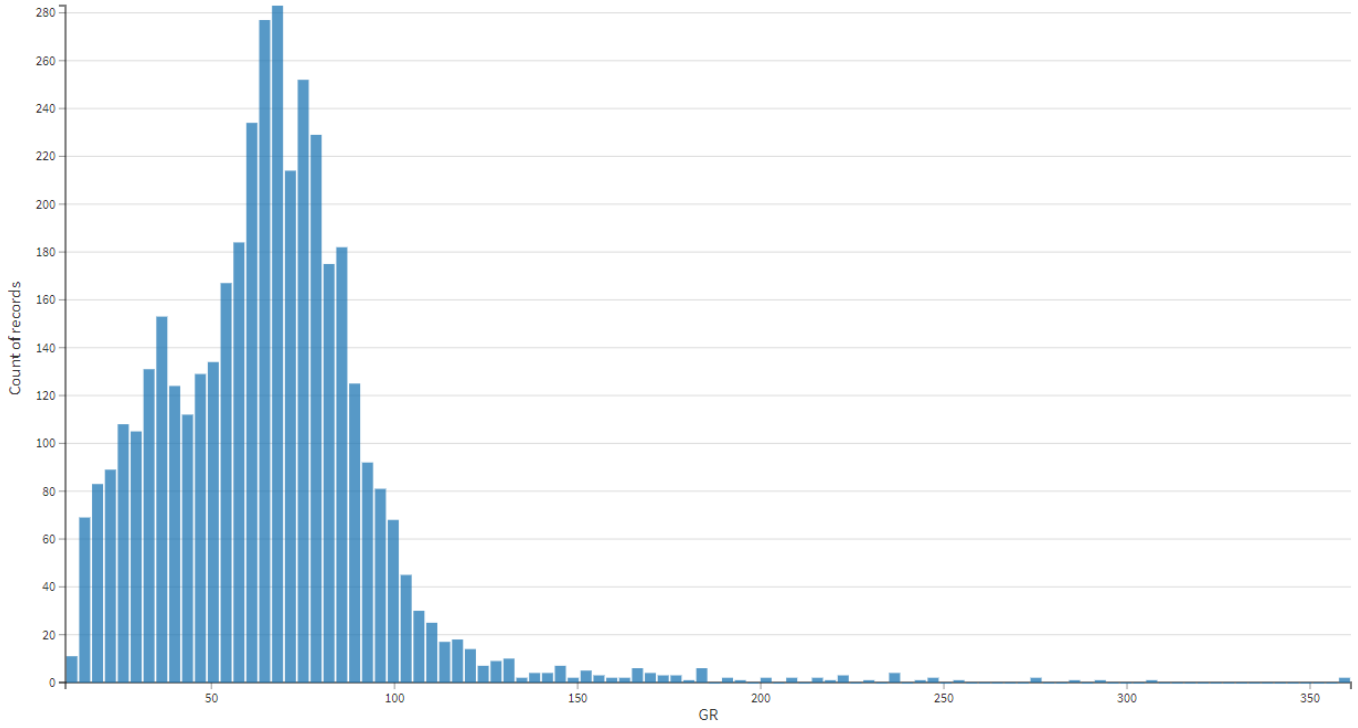
- Most of data cleaning steps can be done by using the "Prepare recipe"

- Every time you add a recipe, Dataiku will create a new dataset

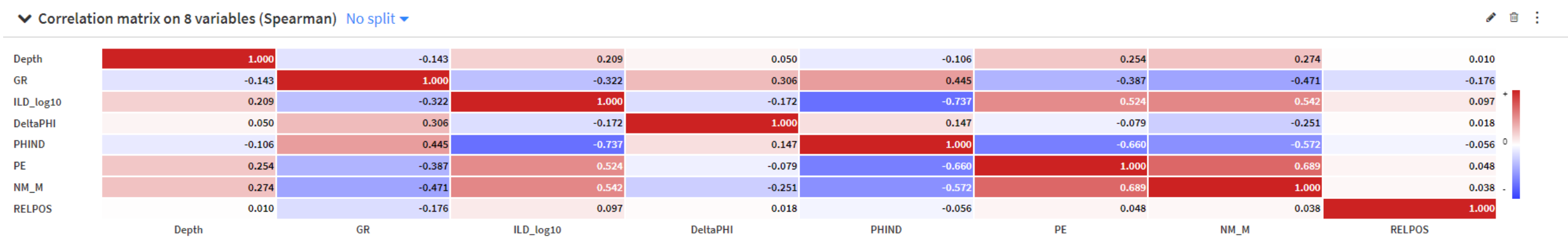
Data exploration

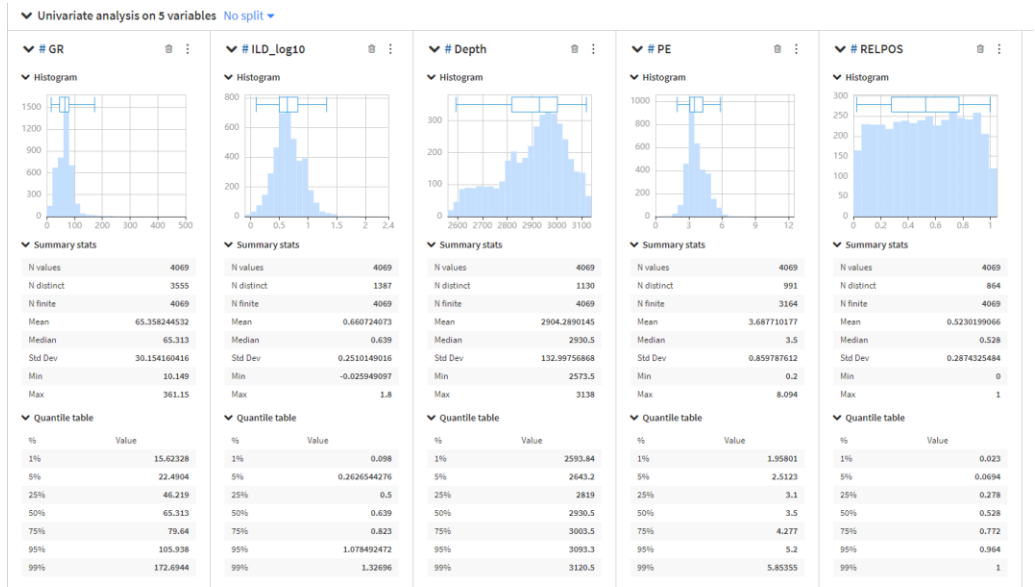


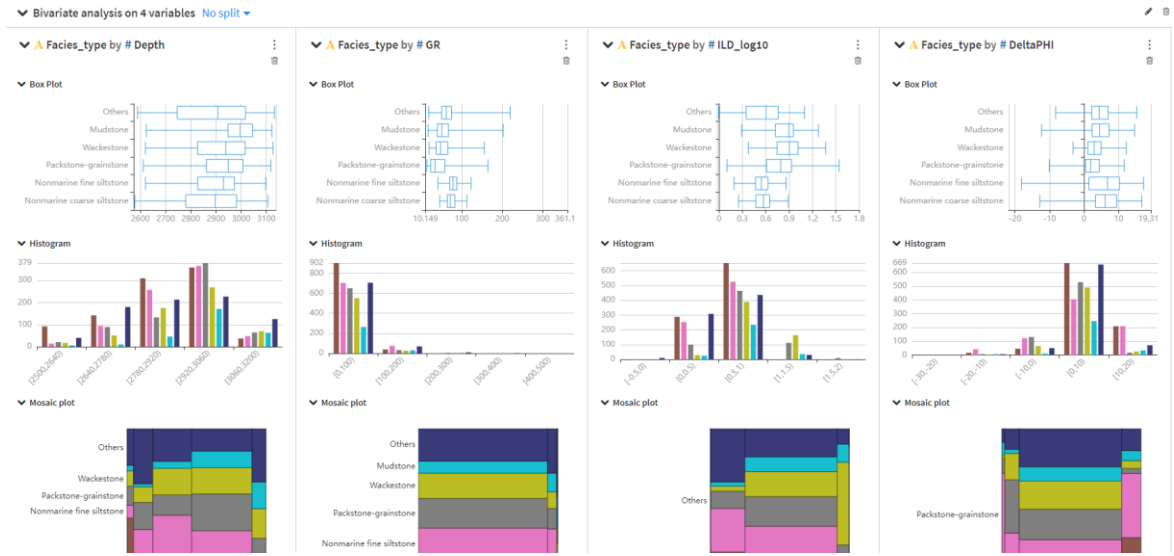




Statistical analysis







Demo Session

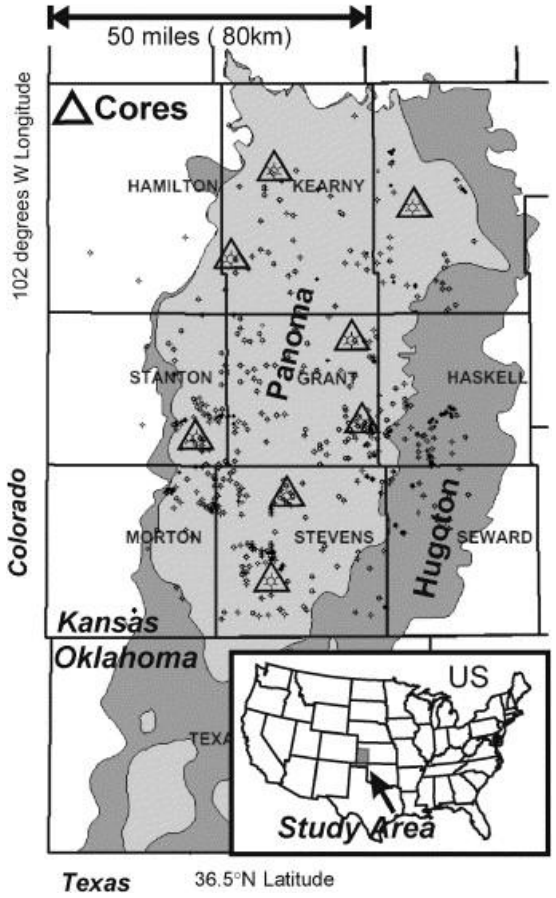
Dataset: Facies classification

Objective: Classify facies

Why?

Increase knowledge of the subsurface

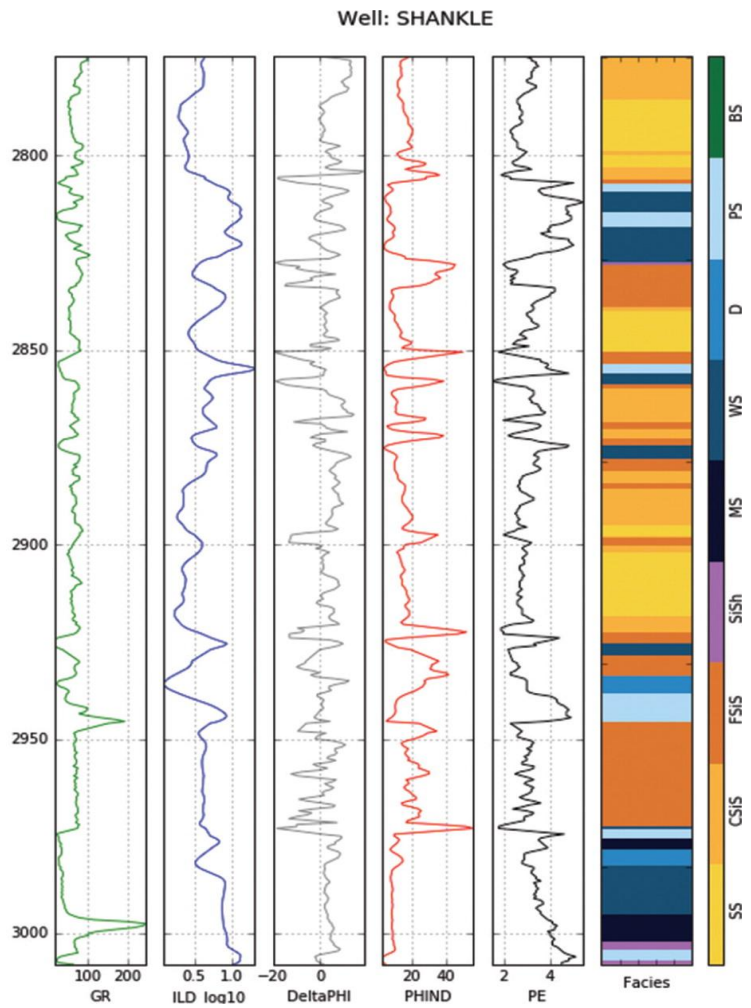
Estimate reservoir capacity



Panoma gas field, in Southwest Kansas, Dubois et al. (2007)

Dataset: Features

- Gamma ray (GR)
- Resistivity (ILD_log10)
- Photoelectric effect (PE)
- Neutron-density porosity difference (DeltaPHI)
- Average neutron-density porosity (PHIND)
- Nonmarine/marine indicator (NM_M)
- Relative position (RELPOS)



Dataset: Labels

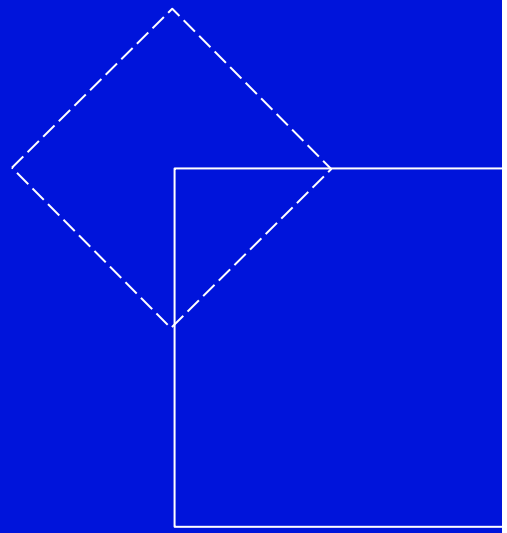
Facies	Description	Label
1	Nonmarine sandstone	SS
2	Nonmarine coarse siltstone	CSiS
3	Nonmarine fine siltstone	FSiS
4	Marine siltstone and shale	SiSh
5	Mudstone	MS
6	Wackestone	WS
7	Dolomite	D
8	Packstone-grainstone	PS
9	Phylloid-algal bafflestone	BS

Large language Models (LLMs)

Abhineet Sinha
Domain Data Scientist-Production

Innovation Factori, Delhi
Gurgaon, India

AGENDA



1. What is Large Language Modeling (LLM)?

2. LLM Overview

3. Hands-On
   a. Summarization and Classification
   b. Retrieval-based question answering

4. What's next?

WHAT IS

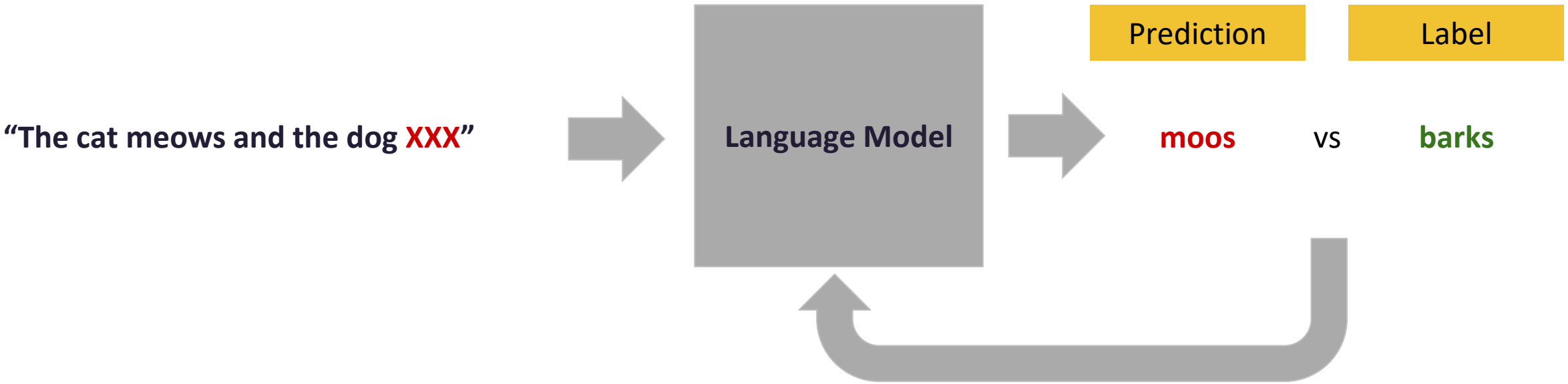
Large Language Modelling

Language model introduction

What is a Language model ?

A Language model is a model that takes text as an input to perform various tasks :

✓ **Classification** ✓ **Translation** ✓ **Text generation / conversation**



Language model introduction

Model training

- LM are trained on each word
- Computational heavy
- Very bad at the beginning and needs a lot of data to get better

“The greatest lesson in life is to ?”

“The greatest lesson in life is to know ?”

“The greatest lesson in life is to know that ?”

“The greatest lesson in life is to know that even ?”

“The greatest lesson in life is to know that even fools ?”

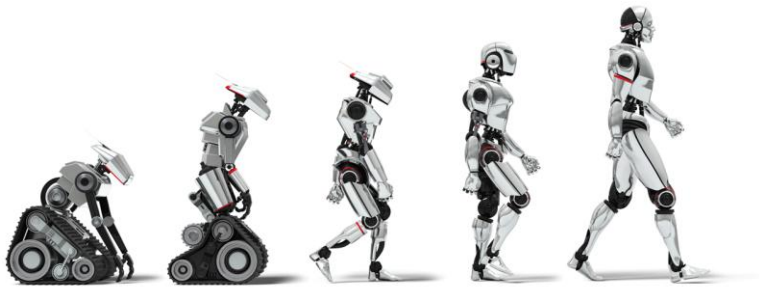
“The greatest lesson in life is to know that even fools are ?”

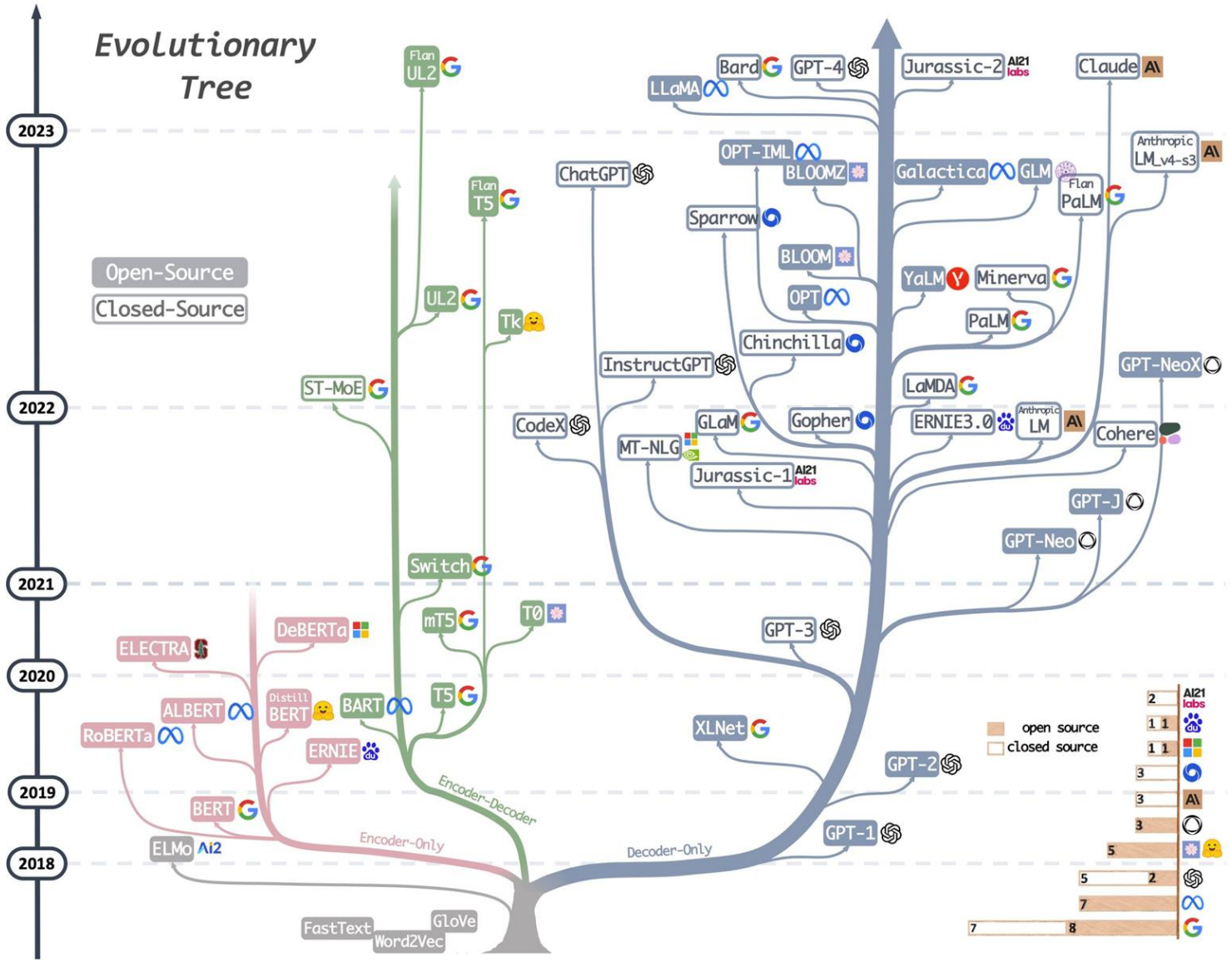
“The greatest lesson in life is to know that even fools are right ?”

“The greatest lesson in life is to know that even fools are right sometimes ?”

Language model introduction

Evolution of LLMs





Language model - quickly explained

Generating Text

- LLMs excel at a specific task: **generating coherent and credible continuations of input texts**, also called prompts.
- LLMs were **trained with colossal amounts of textual data** encompassing diverse sources like books, articles, and websites. Through this process, they acquired remarkable reasoning capabilities which make them much more than writing assistants. In particular, they are especially **effective for a wide range of natural language tasks**.

If you want to build a ship, don't drum up the men to gather wood, divide the work and give orders. Instead, teach them to yearn for the vast and endless sea.

Example of a prompt and a potential LLM response

Language model - quickly explained

Generating Text

- The key to accomplishing natural language tasks with LLMs is to **craft a prompt** in a way that the continuation of this prompt **yields the desired answer**.

- Suppose that we want to classify a product review as positive, negative, or neutral. To do this, we can specify this task in a prompt , insert the text of the product review , and suggest, through the **structure of the prompt, that the immediate next word should be the answer**. When provided with this prompt, an LLM will hopefully provide the correct result

Decide whether the sentiment of the product review is positive, neutral, or negative.

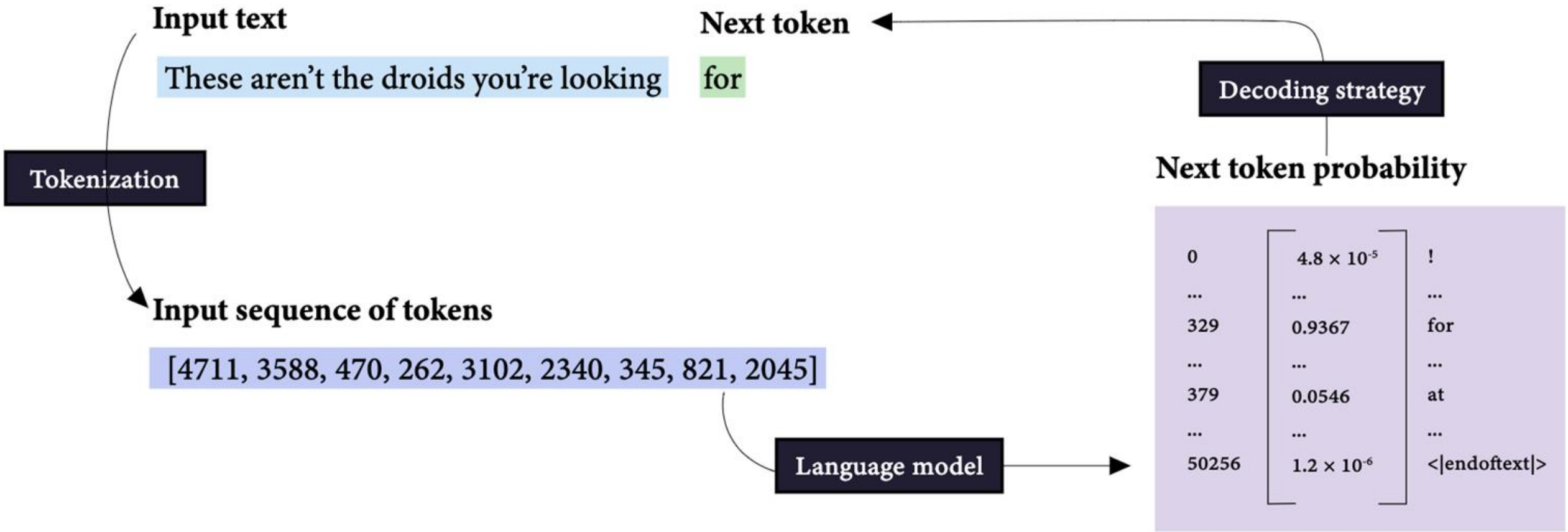
Product review: I never received my package. The customer service was awful

Sentiment: negative

In-context zero-shot learning for a classification task

Language model - quickly explained

Under the hood



Language model - quickly explained

Tokenizing Input Texts and Estimating the Next Token Probability

Tokenizing Input Texts

- Before being processed by a language model, a piece of text is converted into a sequence of tokens which are words or subwords.
- These tokens are selected among a predetermined set of a few tens of thousands of potential tokens and they can simply be represented by integers between 0 and the number of potential tokens minus 1.
- You can test an interactive tool on OpenAI's website if you want to understand more intuitively how tokenization works.

Estimating the Next Token Probability

- A language model takes as an input a sequence of tokens and returns the probability distribution over the next token.
- The length of the input sequence of tokens is limited by the context window size, which is specific to the language model (for example, 4,096 tokens for ChatGPT).

Language model - quickly explained

Implementing a Decoding Strategy

- The probability distribution over the next token can be used to continue the input sequence, one token at a time.

- Several decoding strategies can be used for this. The two simplest approaches are greedy decoding and multinomial sampling (sampling for short).
    - With **greedy decoding**, the next token selected is the one with the highest estimated likelihood.
    - With **sampling**, the next token is randomly chosen based on the estimated probability distribution.
      This estimated probability distribution is based on raw scores computed by the language model and a temperature parameter. By adjusting this parameter, we can make the text generation process more or less deterministic. Temperature 0 corresponds to greedy decoding while high temperatures lead to more original outputs.

How Dataiku can Accelerate the use of LLMs

The problem: barriers impede the full potential of Generative AI

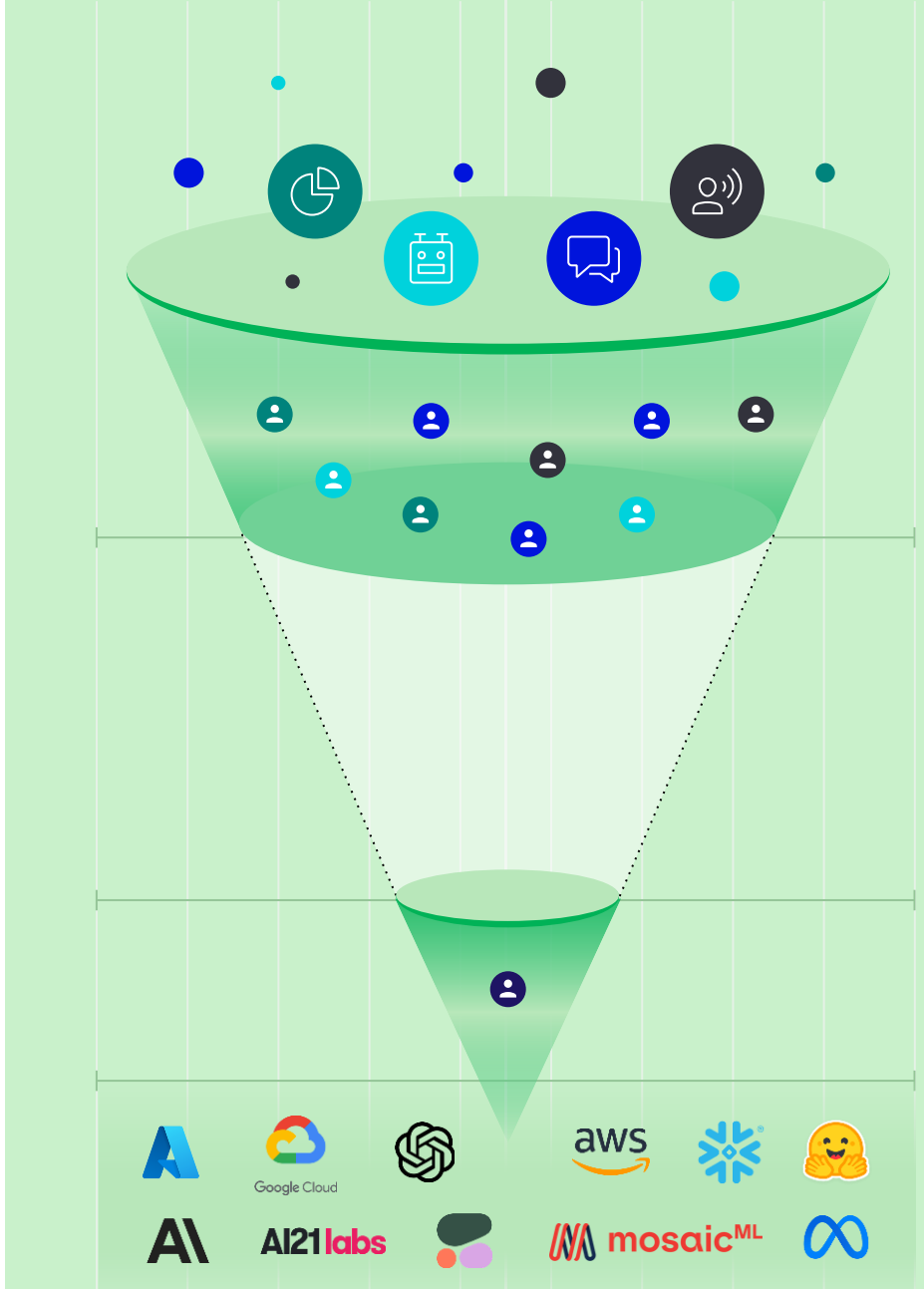
Beyond chatbots, the human-like intelligence of Generative AI could augment and automate thousands of decisions and processes. What's standing in the way?

Not enough builders

Not enough people have the access, tools, and skills to build, validate and leverage GenAI applications

Gap between exciting innovating AI services and enterprise reality

Privacy, Security, and Compliance	**Choice and dependency**	**Cost Control**
Leak of sensitive data or toxic content creation	Gen AI landscape moving very quickly, with a growing choice of models, services and providers.	Creating and running LLMs is expensive.
Need of Governance, control & moderation	Need of Agility to switch from one AI service to another and avoid lock-in	Need to control API fees & infrastructure costs



Dataiku's key differentiators on Gen AI

#1 Ease of use

Simple to advanced techniques accessible through no/low code components

Acceleration for coders & experts

BENEFITS:

Makes Gen AI accessible beyond a narrow pool of LLM experts

Significant speed to value gains for capitalizing on this new tech

#2 Security & Control

A governed, auditable, and cost aware framework with a secure gateway for LLM providers and models to support enterprise LLM applications

BENEFITS:

Reduced risks (security leaks, shared keys, shadow AI)

Cost control & tracking (for chargeback or cost optimization purposes)

#3 Choice & Agility

A level of abstraction that reduces tech dependencies & helps switch between LLM providers without having to rebuild/code projects

BENEFITS:

Flexibility, no lock-in

Future-proof applications can seamlessly take advantage of new tech as it emerges

#4 Holistic Platform

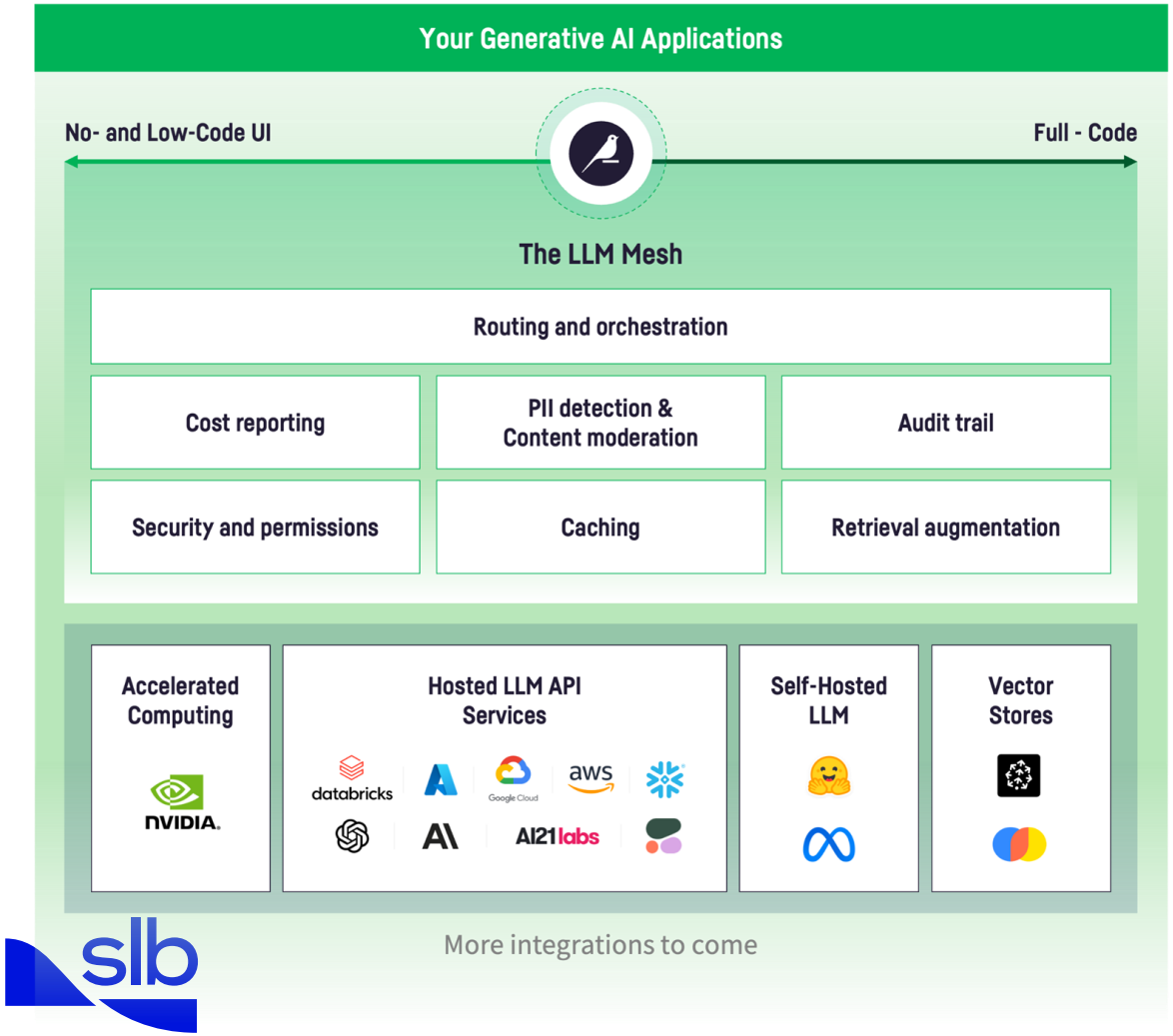
Extend and amplify existing analytic and predictive workflows with Gen AI. Dataiku's platform can meet all your analytics & AI needs, from end-to-end.

BENEFITS:

Amplify value of Gen AI by integrating it in core analytics and ML workflows

No need to buy siloed point solutions for Gen AI

The LLM Mesh: The Common Backbone for All GenAI Applications



Decouple the application and AI service layers

Enforce a secure gateway to approved LLM services and maintain an audit trail

Ensure safe use by defining filters for queries and responses such as PII detection, content moderation...

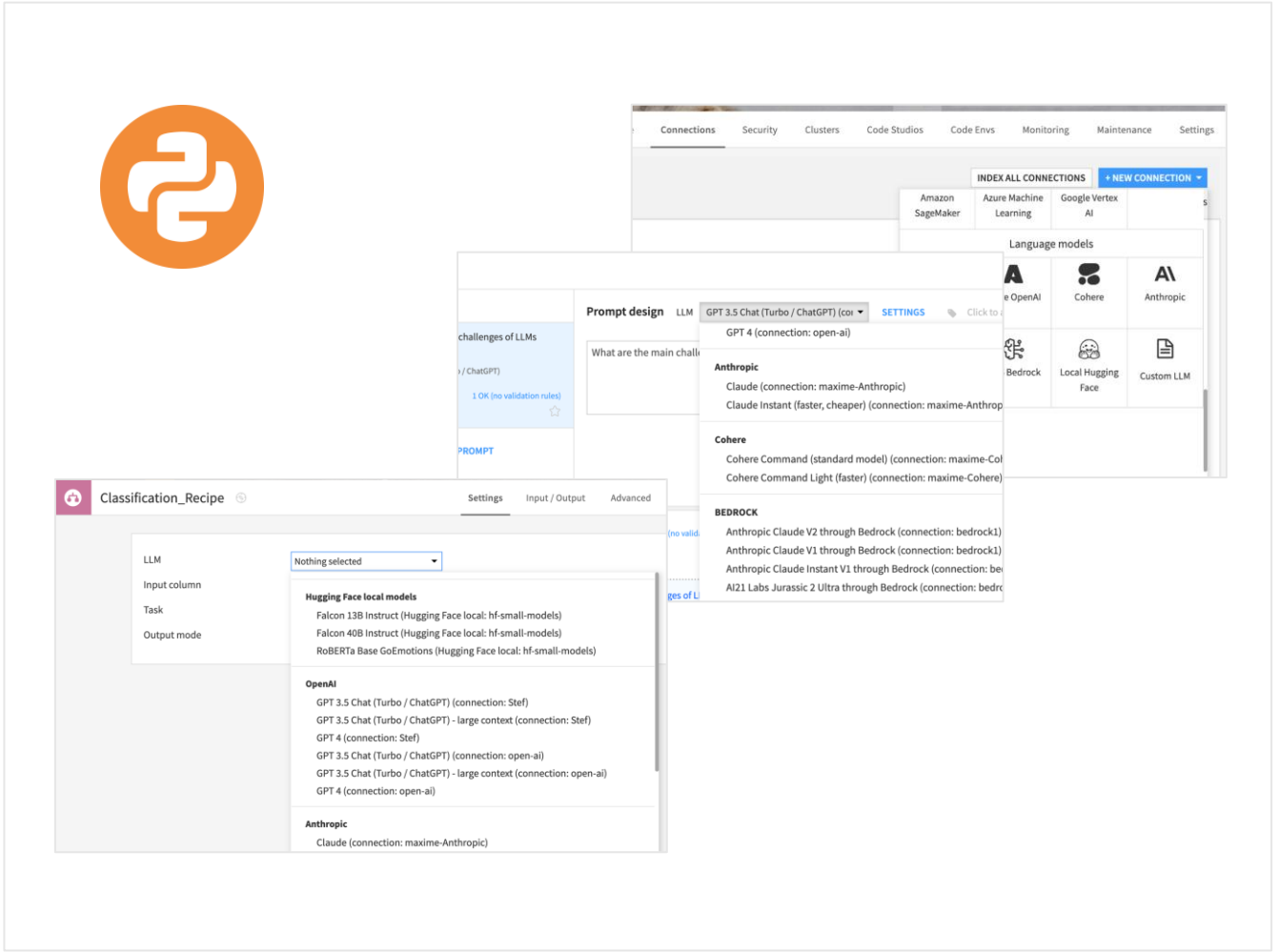
Control costs and avoid unnecessary re-generation with caching

Enrich queries and responses with built-in retrieval augmentation

LLM Mesh: Decoupling the application and AI service layers

Maintain choice and agility as new technologies emerge

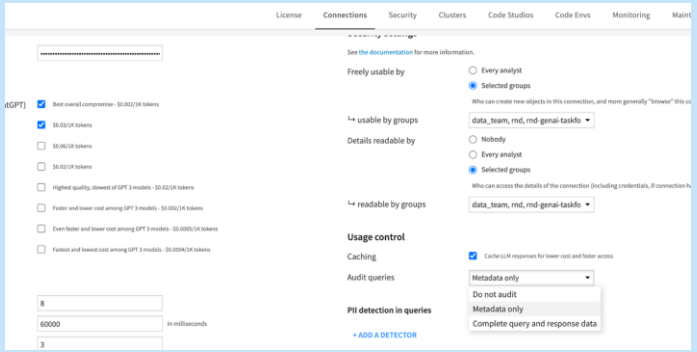
- **Connect to the AI services of your choice including LLMs and Vector DB**
  - LLM & Vector DB connections
- AI services and their various models are available throughout the Dataiku platform (visual NLP recipes, Prompt Studios, code recipes, notebooks, IDE, …)
- Keep your choices open by making your applications agnostic to underlying models and providers
- Test different services to balance cost and performance

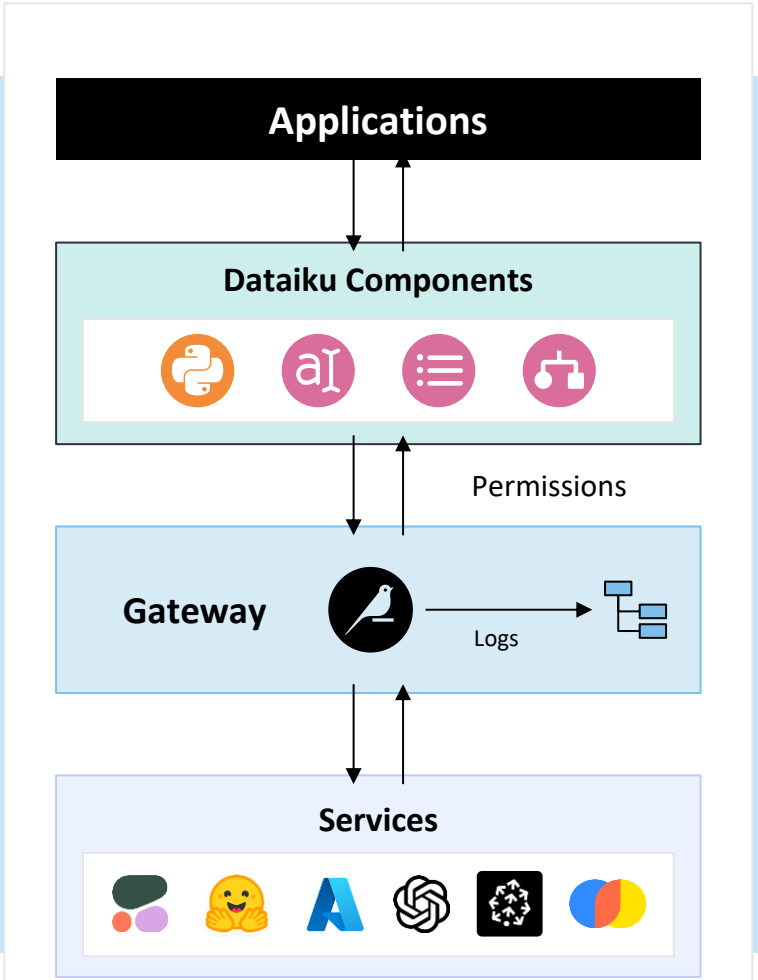


LLM Mesh: Decoupling the application and AI service layers

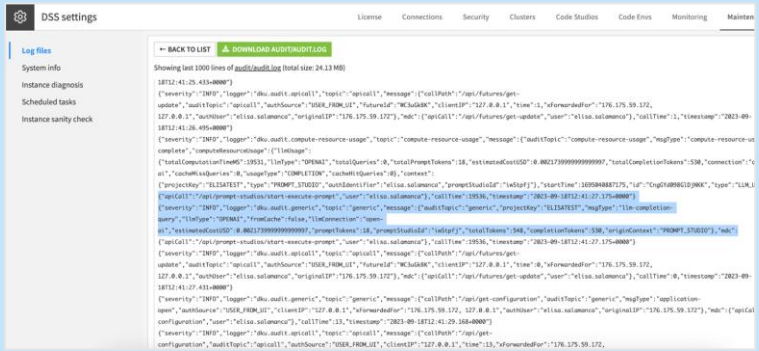
Maintain choice and agility as new technologies emerge

Centrally manage access keys to different services to ensure that only approved services are used





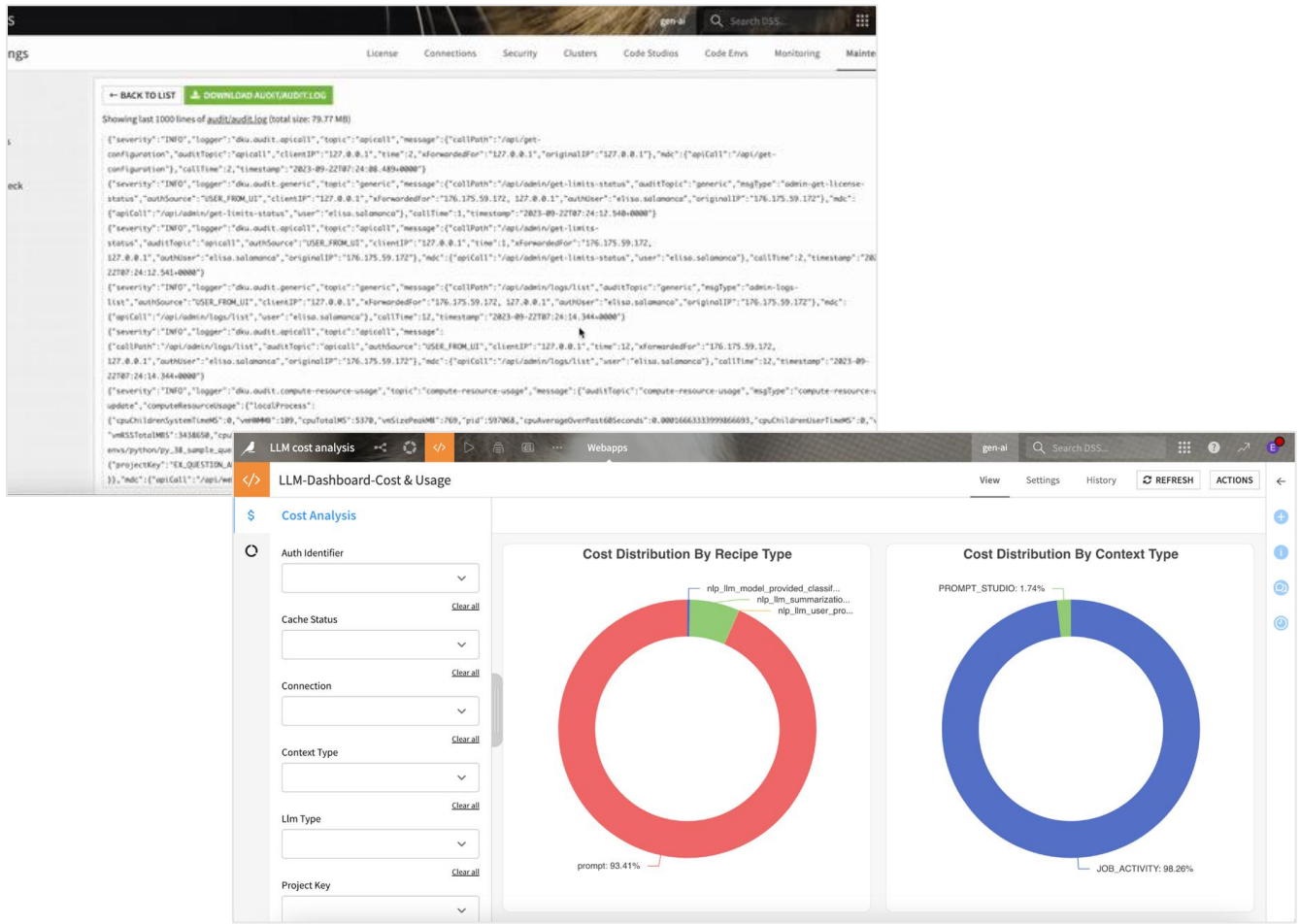
Log all queries and requests to all services, creating a secure, searchable trail of all LLM use in the enterprise



LLM Mesh: Control and optimize costs

Monitor costs and cache common responses to avoid re-generation

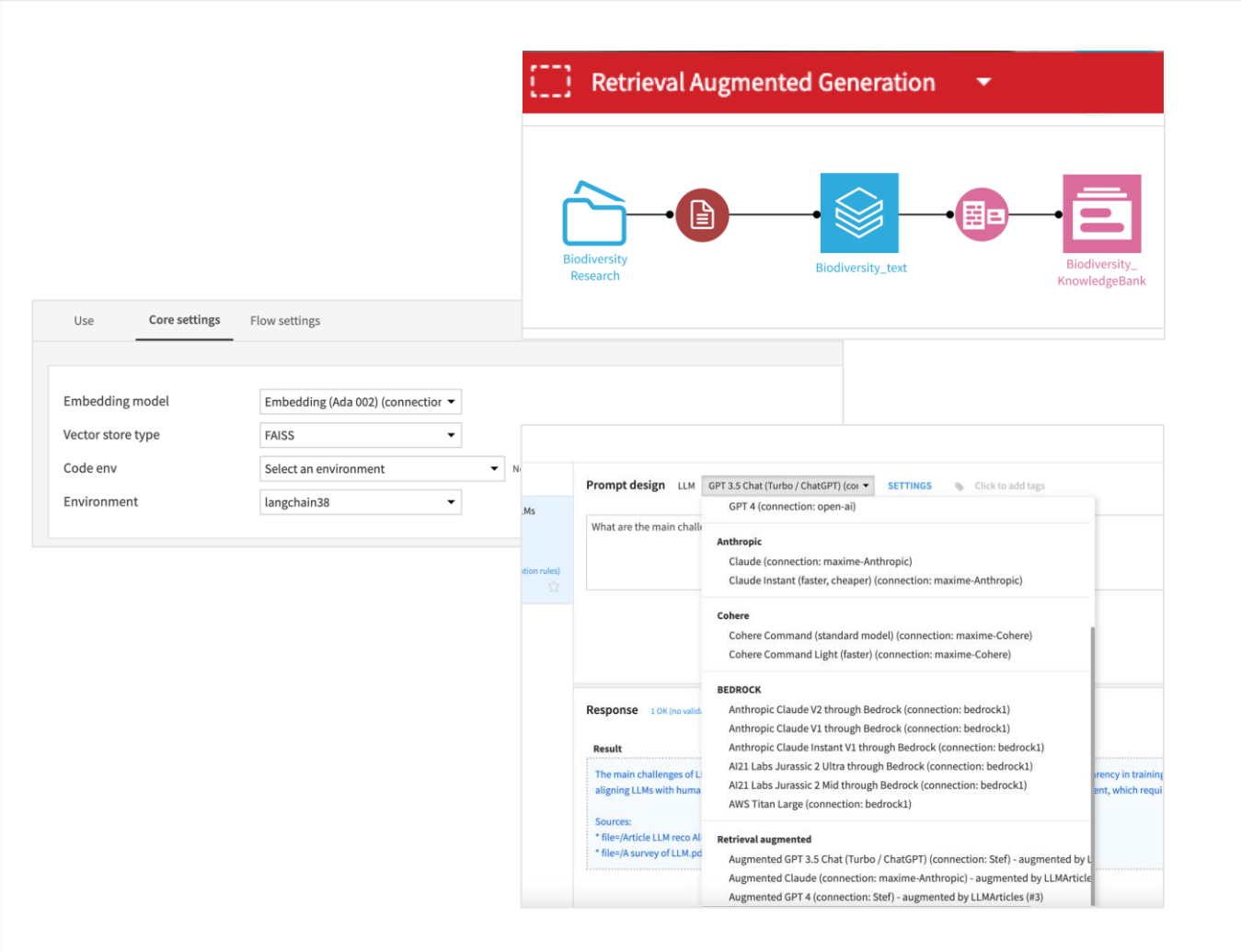
- Cache responses to common queries to avoid unnecessary re-generation and associated costs
- Use Compute Resources Usage in Dataiku to generate fine-grained cost reports



LLM Mesh: Built-in retrieval augmentation

Enrich queries and responses with internal documentation

- Work from documents in any directory, using built-in pre-processing (OCR, translation, etc…) if needed
- Select the model of your choice to create the embeddings
- Easily create knowledge banks using the vector store of your choice
- Use the augmented model elsewhere in Dataiku (for example, Prompt Studios) to build knowledge-enriched apps



Case 1: Use the Capabilities LLMs for Common Language Tasks

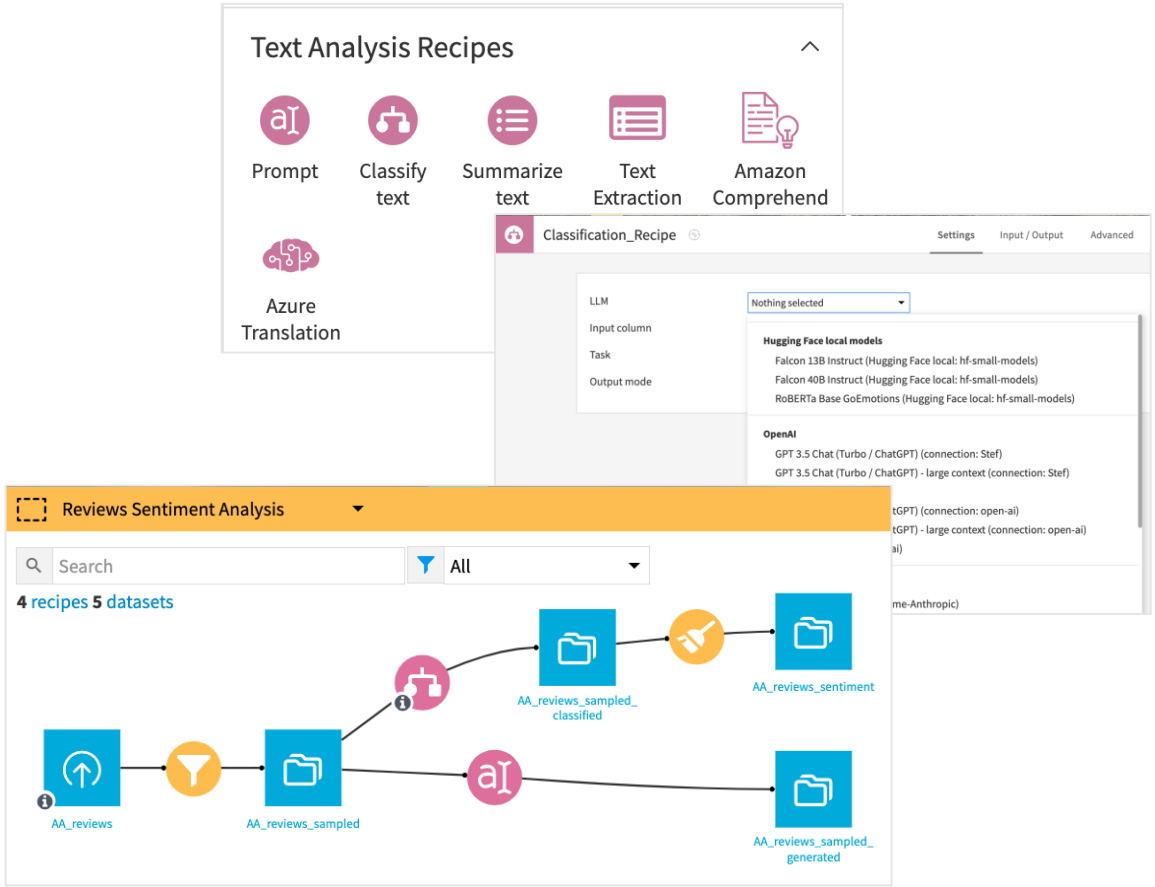
Off-the-shelf, visual components bring the capabilities of LLM to the entire enterprise

With Dataiku, deliver powerful **no-code** LLM-powered projects and pipelines

Flexible, **provider-independent approach:** Leverage any provider's services OR maintain total privacy with local models

Leverage Dataiku's growing library of ready-to-use **visual LLM-powered recipes:**

- Entity extraction
- Text classification
- Text summarization
- Translation
- And more

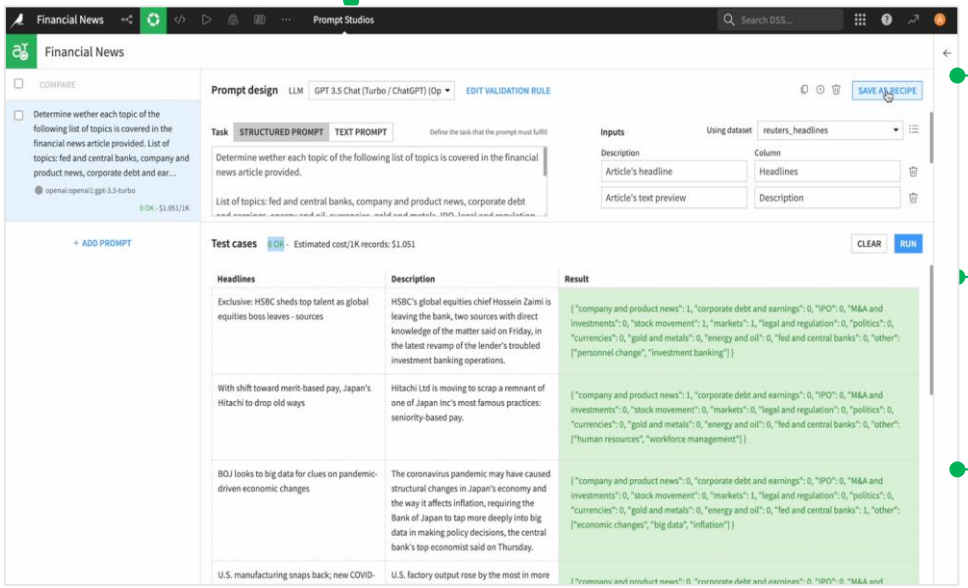


Case 2: Prompt Engineering

Customize LLM outputs by refining prompt context, format, length, and instructions

- Prompt Studios in Dataiku help you design, evaluate, and operationalize optimized LLM prompts in a no-code way
- Model-agnostic so you can compare results across different models & providers to identify the best model to achieve your specific goals
- Built-in cost estimates inform important trade-off decisions between cost & performance during the design phase
- Deploy your prompt in your data pipeline

Test different models, including cloud, local or your fine-tuned models



Save as **Prompt Recipe** in your data pipeline

Estimate and compare running costs

Validate model responses

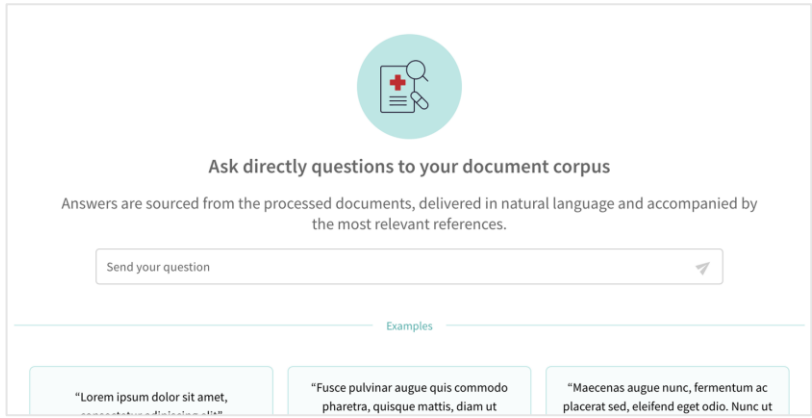
Leverage a library of templates, promoting reuse and best practices.

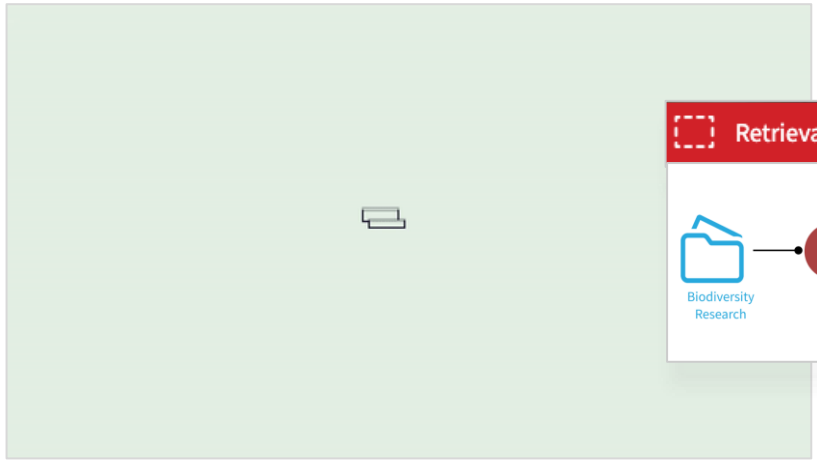
Case 3: Knowledge Bank - Add Your Internal Knowledge

Incorporate information retrieval and LLM augmentation methods to build Q&A apps & chatbots

- Develop and deliver Q&A-type LLM applications backed by your own proprietary documents in a no-code way

- Leverage built-in components for Retrieval Augmented Generation (RAG) and semantic search :
  - Text extraction from files and OCR
  - Embedding Recipe
  - Knowledge banks (using vector stores)

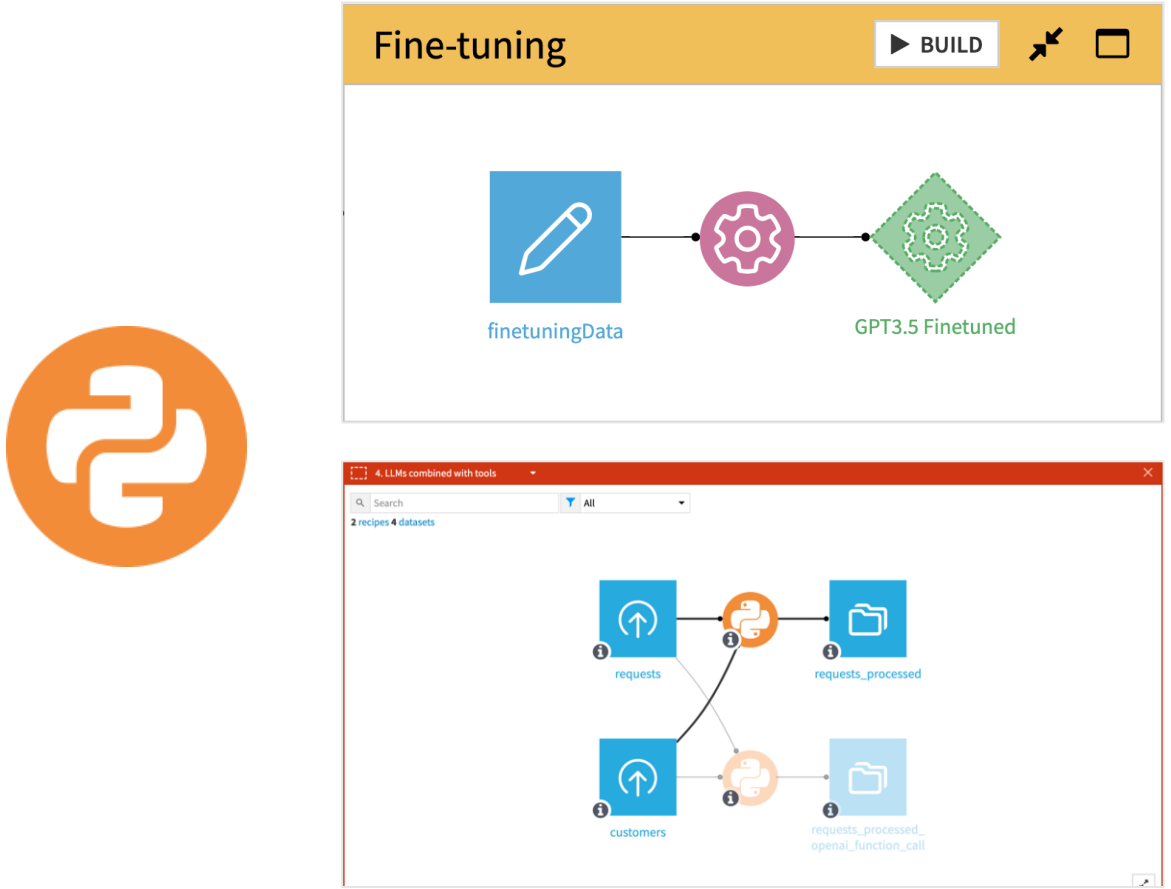




Case 4: Advanced Techniques

Incorporate LLM agents, set up complex reasoning chains, and fine-tune models

- **LLM agents:** Orchestrate complex reasoning tasks and chains of prompts, data sources, and logic with your own code, and leverage our integration with Langchain API
- Leverage LLM providers fine-tuning services in a seamless way with our fine-tuning recipe
- Fine-tune or pre-train your own models through code



Hands-On Session

Thank You