JOIN LIVE

TOP 3 MACHINE LEARNING PROBLEM AND THEIR SOLUTION

HOW TO WORK WITH LARGE DATASET? (WITH MULTI MILLION ROWS)
HOW TO MANAGE CHANGE IN DATA?
HOW TO MANAGE CHANGE IN ML MODEL (BOTH IN DEVELOPMENT AND PRODUCTION)?

LIVE ONLINE SESSION

ON 06-AUG-2022@ 7.00 - 08.30 PM IST (SATURDAY)
@ 9.30 - 11.00 AM EST
@ 2.30 - 04.00 PM LONDON

JOIN FREE HTTPS://FORMS.GLE/7BEFLRXK97F2FKWB7

BY PRASANNA VENKATESH J

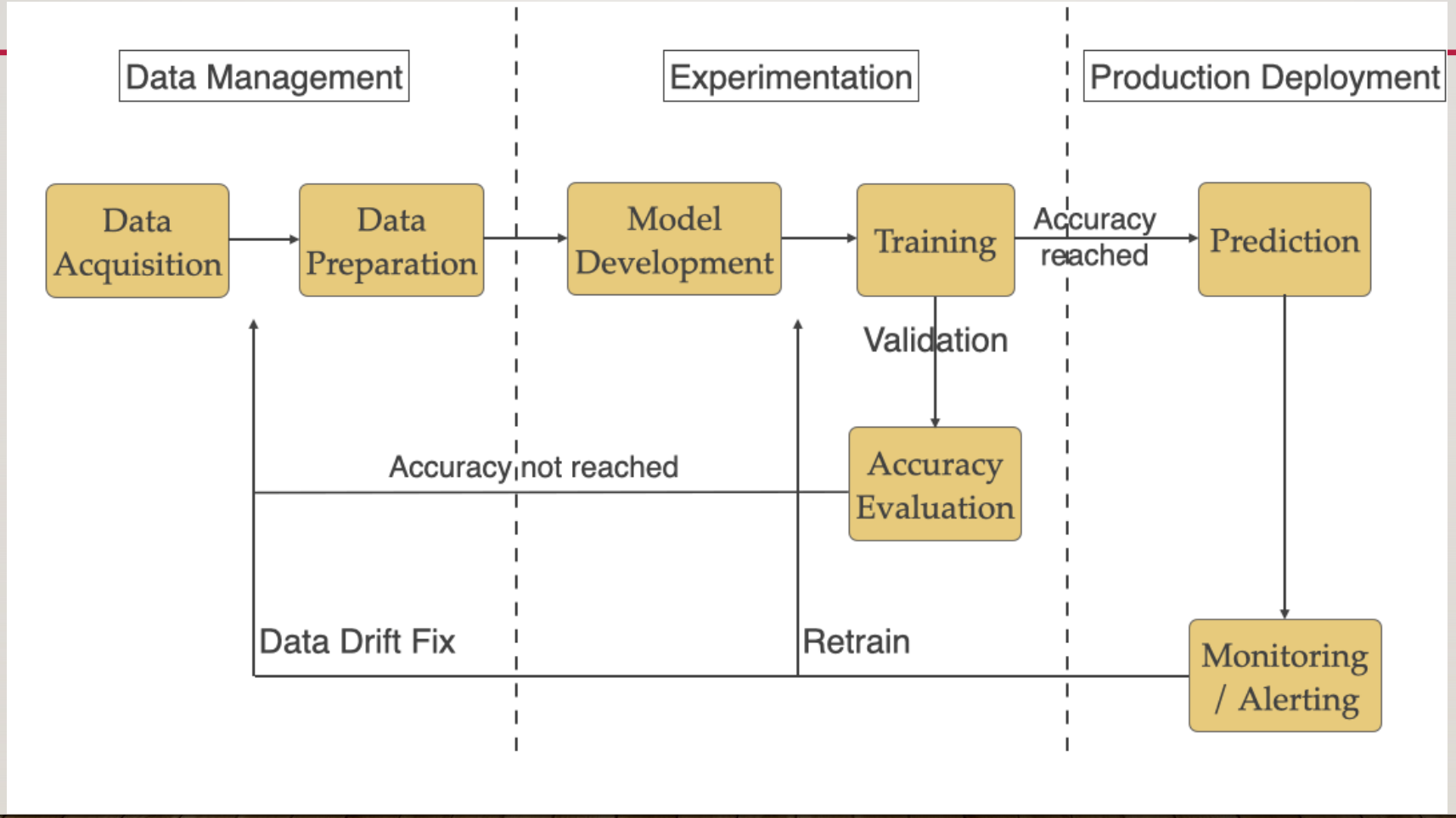
AI & CLOUD EVANGELIST

AGENDA

- Top3 problem's in Machine Learning ntroduction:
  - Problem-1: Load large data
    - Approach and solutions.
    - Sample Demo
  - Manage data versioning
    - Various options and solutions
    - Sample Demo
  - Manage production model
    - Deploying and managing solution in live
    - Changes and Sample Demo.

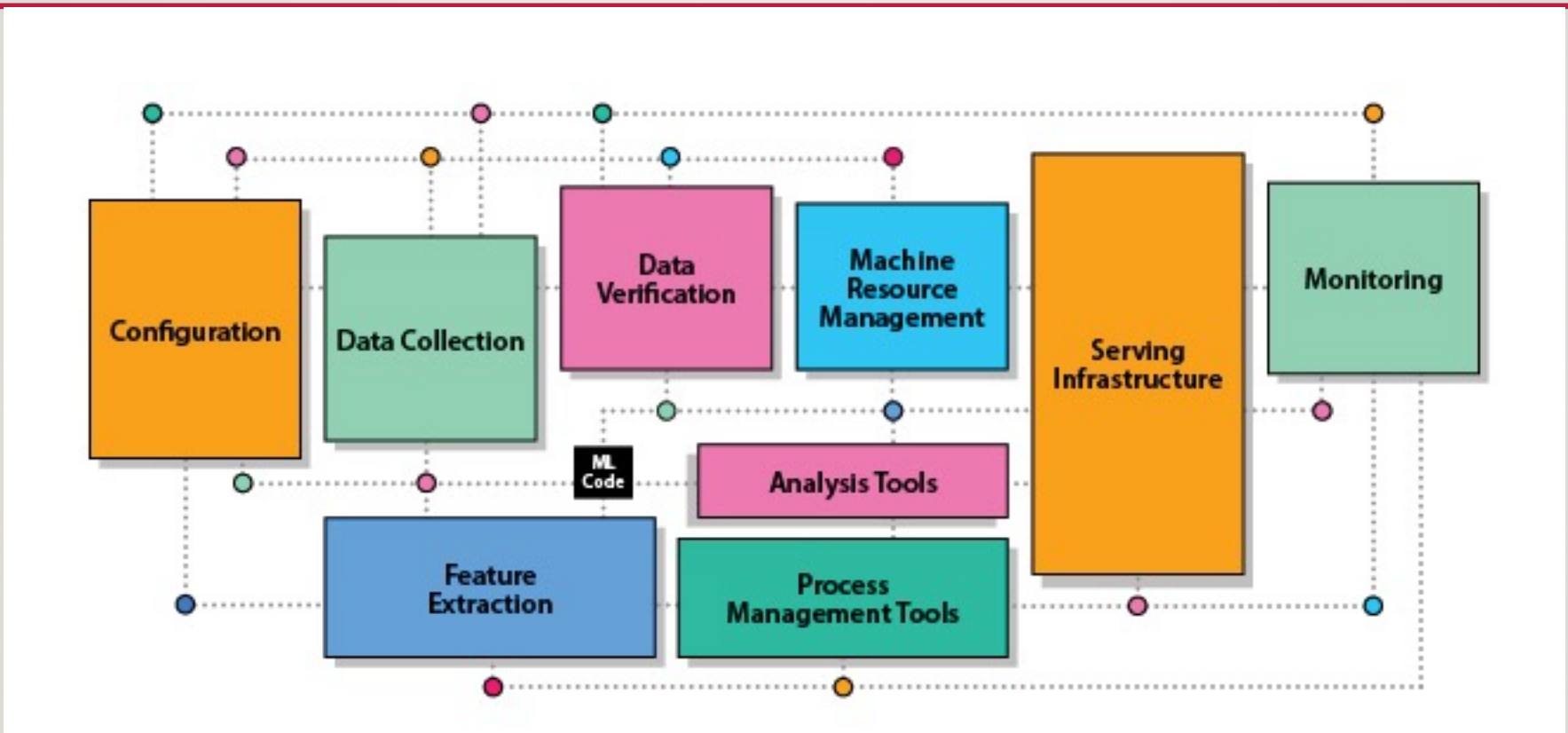
prasannavj@gmail.com

MACHINE LEARNING WORKFLOW



prasannavj@gmail.com

MACHINE LEARNING WORKFLOW



prasannavj@gmail.com

PROBLEM-1

- Load large data set and work with large data set.
- CSV is not a good option what alternate to use?
- If my data source is in CSV with large memory footprint, how do I use it for my ML training?
- My training resources are limited with only one GPU?
- How to create quick baseline model outcome?

prasannavj@gmail.com

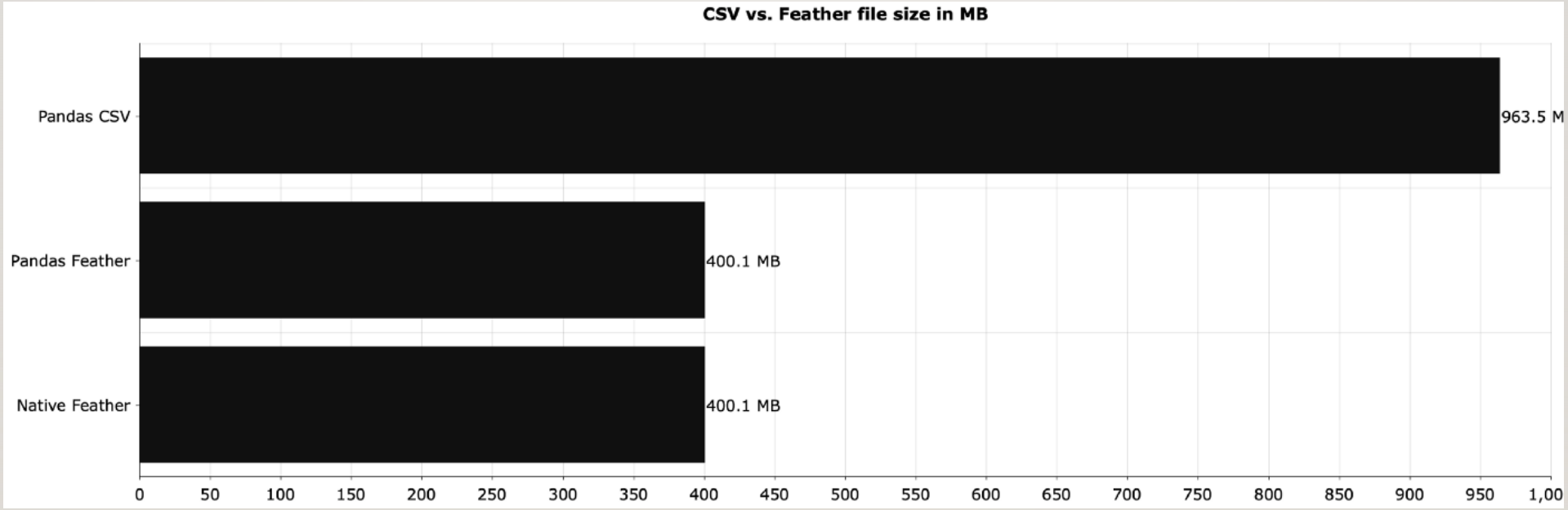
ALTERNATE TO CSV FORMAT

Alternate dataformat to work with large data effectively.

- Feather
- Parquet
- Jay

FEATHER DATA FORMAT

- Comparing to CSV vs Feather size comparison of same data.

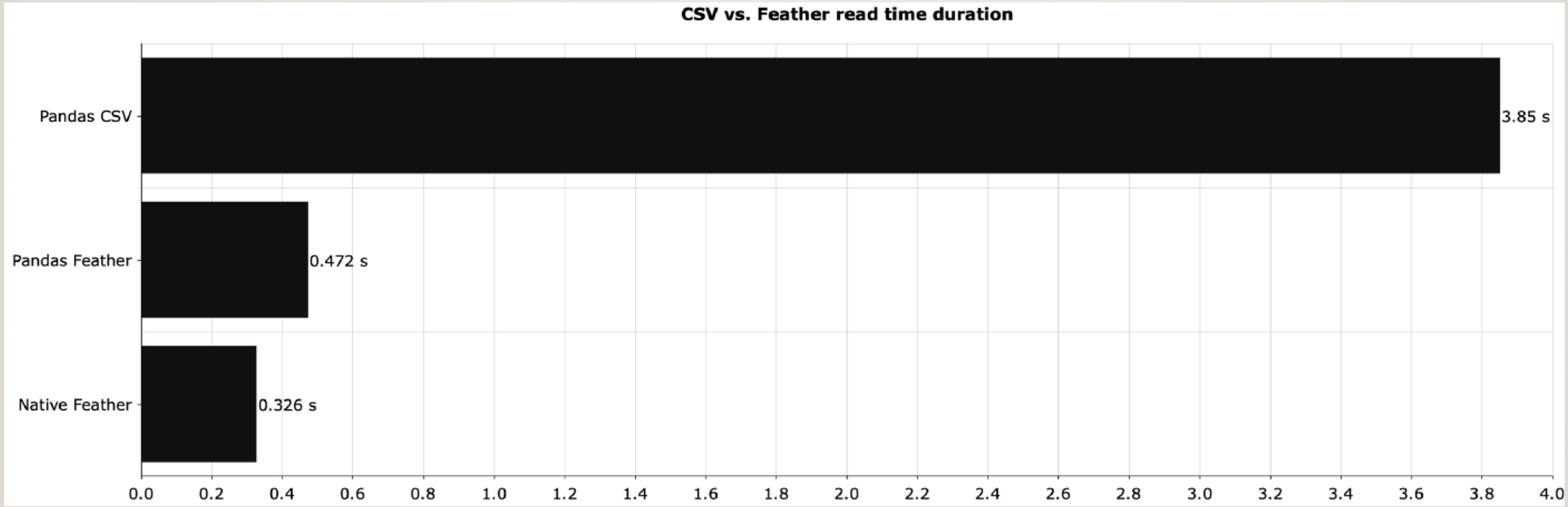


https://towardsdatascience.com/stop-using-csvs-for-storage-this-file-format-is-150-times-faster-158bd322074e

prasannavj@gmail.com

FEATHER DATA FORMAT

- CSV vs Feather Read time Duration

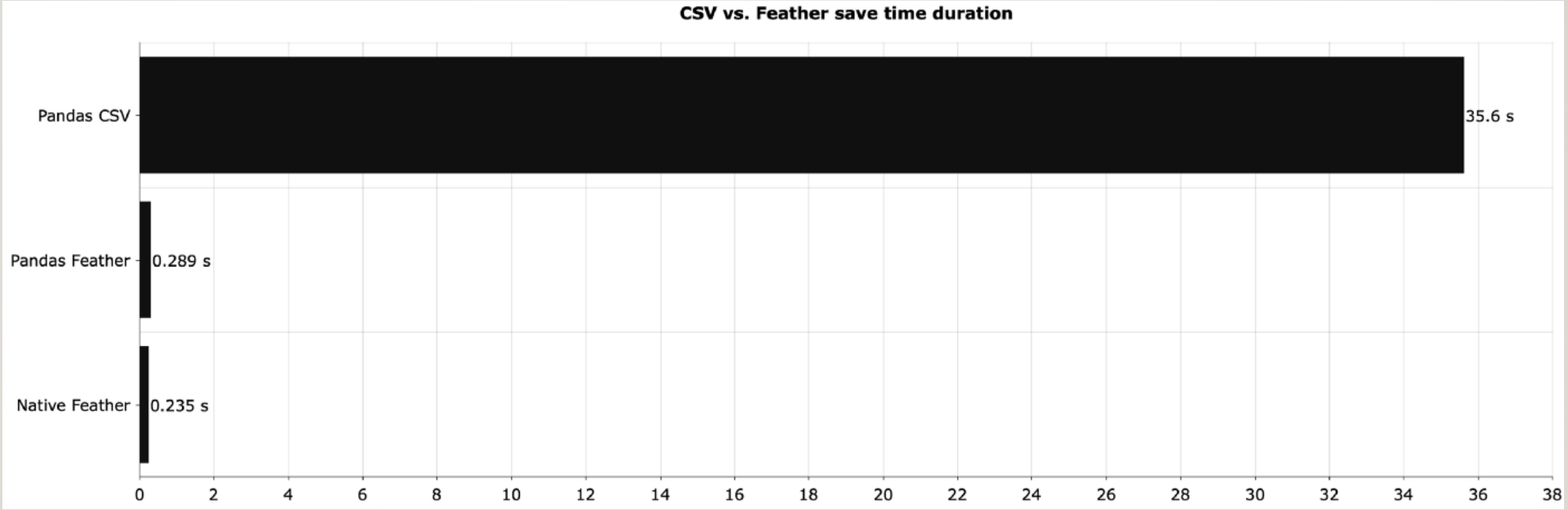


https://towardsdatascience.com/stop-using-csvs-for-storage-this-file-format-is-150-times-faster-158bd322074e

prasannavj@gmail.com

FEATHER DATA FORMAT

- CSV vs Feather Save Time Duration



https://towardsdatascience.com/stop-using-csvs-for-storage-this-file-format-is-150-times-faster-158bd322074e

SUPPORT FOR PANDA INTEGRATION

- Usage Sample (Works only with Python >3.8)
- *Install Feather*
  - *Pip Install Feather*
- Convert from CSV to Feather
  - Load CSV as Dataframe
  - Dataframe to Feather example
    - df.loc[2000000:4000000].reset_index().to_feather("./myfeather.ftr")
    - df.loc[2000000:4000000].reset_index().to_feather("./myfeather.feather")

prasannavj@gmail.com

APACHE PARQUE FILE FORMET

- Apache Parquet is an open source, column-oriented data file format designed for efficient data storage and retrieval.
- It is similar to other column-oriented data format like Hadoop.
- Support for working with large data size and multiple data formats.

prasannavj@gmail.com

CHARACTERISTICS OF PARQUET

- **Free and open-source file format.**
- **Language agnostic.**
- **Column-based format -** files are organized by column, rather than by row, which saves storage space and speeds up analytics queries.
- **Highly efficient** data compression and decompression.
- **Supports complex data types** and advanced nested data structures.

prasannavj@gmail.com

BENEFIT OF PARQUET

- **Good for storing big data of any kind** (structured data tables, images, videos, documents).
- **Saves on cloud storage space** by using highly efficient column-wise compression, and flexible encoding schemes for columns with different data types.
- **Increased data throughput and performance** using techniques like data skipping, whereby queries that fetch specific column values need not read the entire row of data.

prasannavj@gmail.com

LIVE DEMO

- Parquest storing and retreaving data.
- Converting the data to CSV to Parquet format.

- Git Hub link: https://github.com/prasannavj/MACHINELEARNINGPROBLEMS

PROBLEM -2 – MANAGING THE CHANGE IN VERSIONS OF MODEL AND MANAGE MODEL DEVELOPMENT

- I want to build my code locally but run it on cloud?
- I want to collaborate with fellow data scientist to work on new algorithms?
- I want to track how the model development evolves.
- I am running a model with multiple epoch and my model failed in between? Can I run from where I stopped?
- I saw my previous model was performing better than new one? Can I go back to old model?

TECHNOLOGY TO HELP EFFECTIVE MODEL DEVELOPMENT

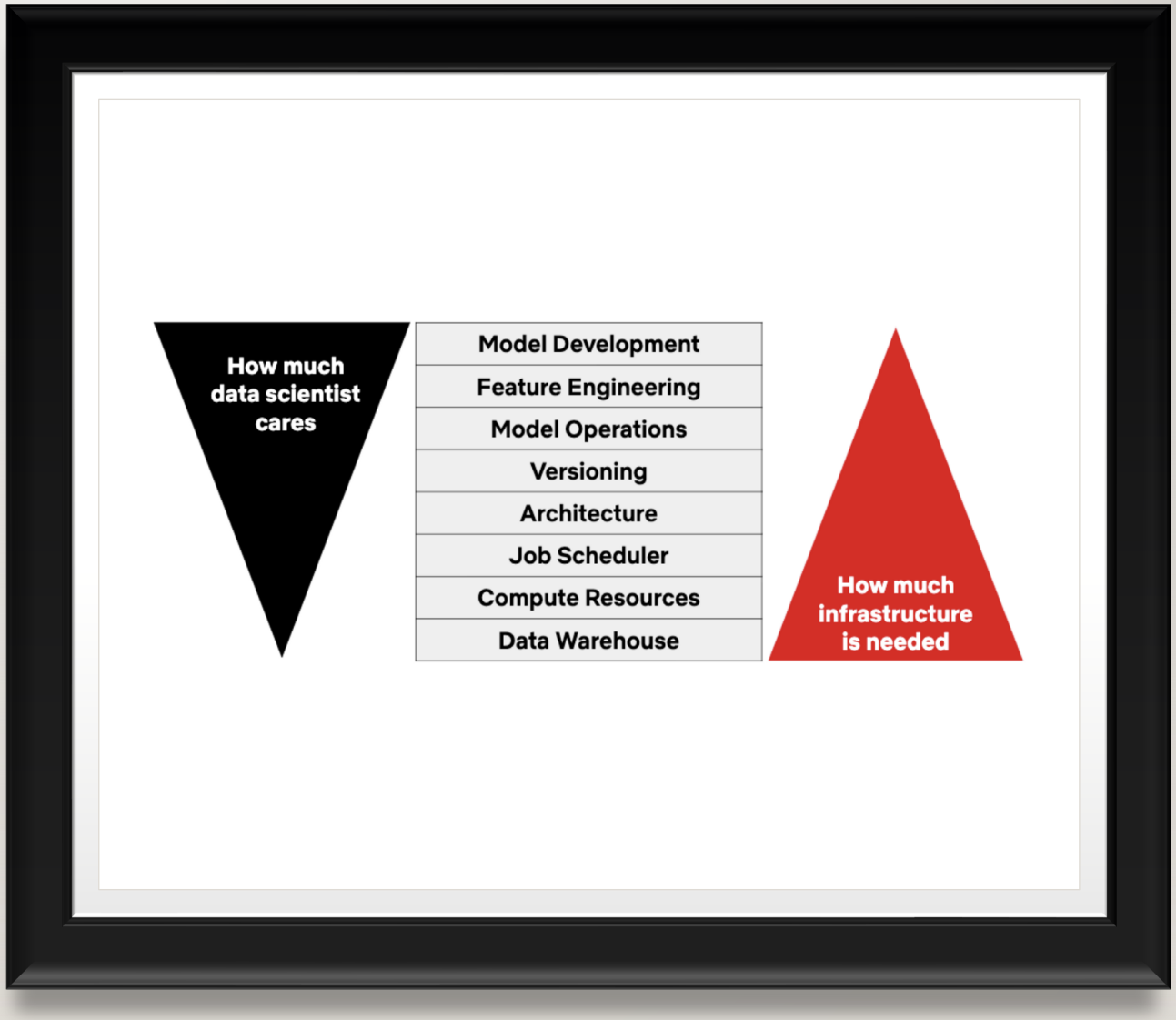
Open Source Platforms

Properatory Platforms

prasannavj@gmail.com

METAFLOW

- Meta flow snapshots your code, data, and dependencies automatically in a content-addressed datastore (like local machine or S3 Buckets in cloud).



FEATURES OF IMPORTANCE

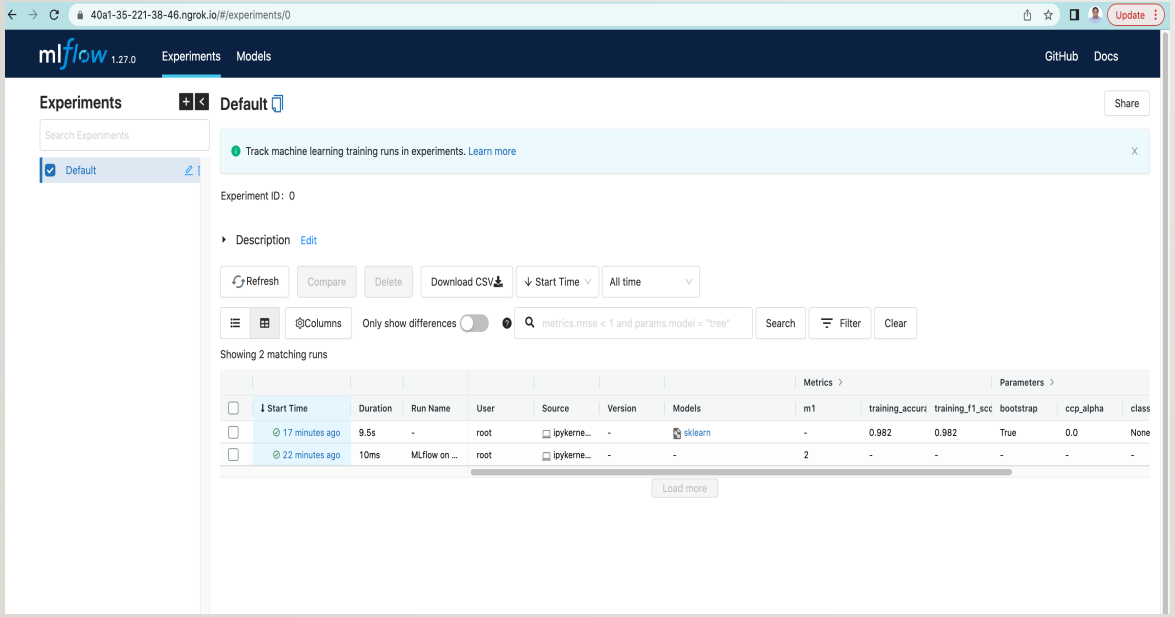
- **Collaboration with fellow data scientist**
- **Resuming a run**
- **Hybrid run (Local and Cloud Run)**
- **Inspecting Run Metadata**
- **Multiple version of same metadata**

METAFLOW INTRODUCTION

- Open source Framework for creating and executing data science workflows and comes equipped with built-in features to:
  - manage compute resources
  - perform containerized runs
  - manage external dependencies
  - version, replay and resume workflow runs
  - Client API to inspect past runs suited for notebooks
  - move back and forth between local (e.g. on a laptop) vs remote (on the cloud) modes of execution

prasannavj@gmail.com

MLFLOW INTRODUCTION



- MLflow is a great open-source tool
- Track your model runs, including model parameters, metrics, results, data used, and your code.
- deploying models, packaging your code for reproducibility, and storing your models

prasannavj@gmail.com

MLFLOW FOR MODEL EVALUATION

- **mlflow.start_run()** : It is used to start a new ***run (*** *a session* ***)*** within an ***experiment (*** *collection of* ***runs)***.
- **mlflow.log_param(key, value)** : It is used to log a key-value pair. key being a ***hyperparameter*** name or any parameter's name.
- **mlflow.sklearn.autolog()** : It automatically logs ***hyperparameters***
- **mlflow.sklearn.eval_and_log_metrics(model, X_test, y_test, prefix='val_') :** It evaluates the model on the test set and logs the ***computed metrics***.
- **mlflow.sklearn.load_model()** : It loads a saved model, given the ***run_id (****id given to a* ***run)*** and the name with which the model was saved.
- **mlflow.end_run()** : It is used to end a current ***run.***

prasannavj@gmail.com

MLFLOW LIVE DEMO

- Source code for sample Mlflow Live demo – Github link
- https://github.com/prasannavj/MACHINELEARNINGPROBLEMS

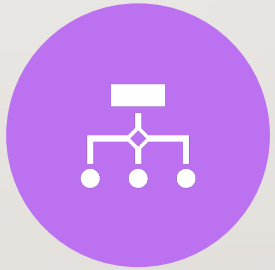
PROBLEM 3 - INTRODUCTION

- How do I track my model performance in production?
- Is data changed in production? What is the current performance of my model in Live?

prasannavj@gmail.com

MONITORING FOR CHANGE IN DATA

DATA DRIFT



Change in Model performance once the Model is Live.

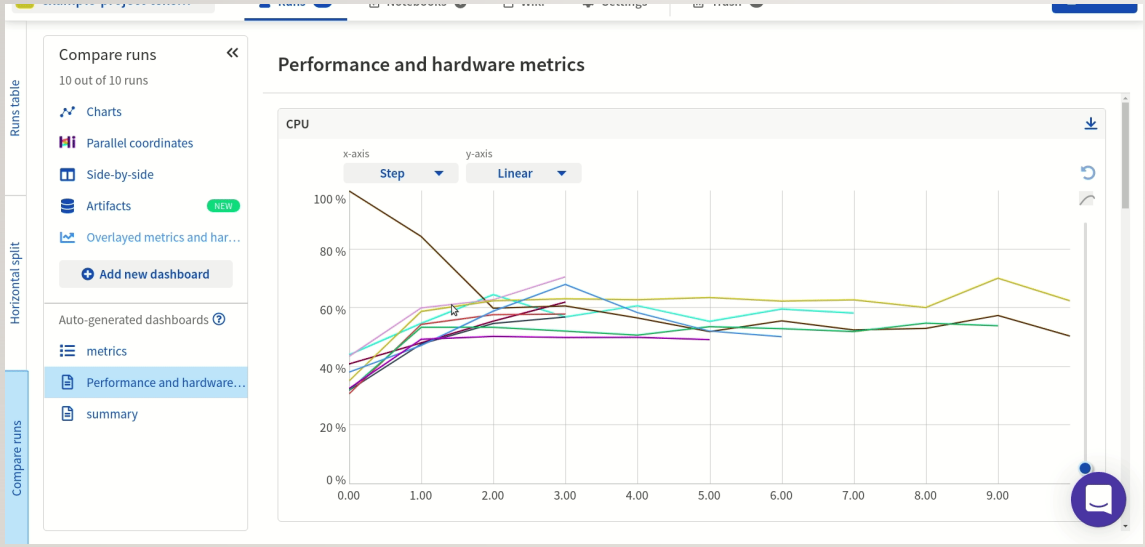
Continues monitor and evaluate how the Model performs

LIVE ML Monitoring to help to understand change in Production ML Model

prasannavj@gmail.com

TOOLS AND TECH FOR LIVE ML MONITORING

Neptune.AI



Grafana + Prometheus

WHYLABS

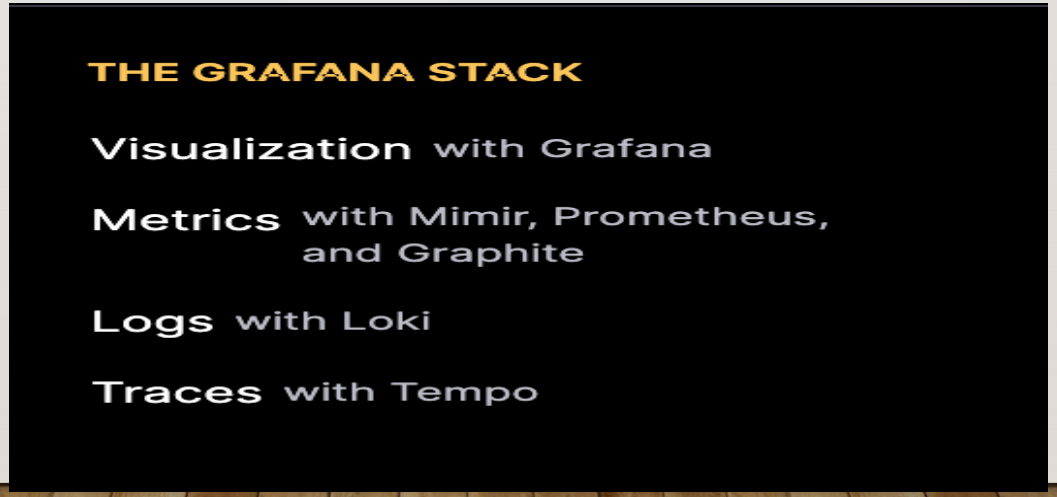
EVIDENTLY AI

arize

prasannavj@gmail.com

GRAFANA + PROMETHEUS – INTRODUCTION

- Open source
- Query, visualize, alert on, and explore your metrics, logs, and traces wherever they are stored.
- Convert a Time Series data to a wonderful visualization to enjoy the outcome.
- Prometheus Gates



prasannavj@gmail.com

GRAFANA + PROMETHEUS

- For Production ML monitoring solution.
- Live Demo
  - Build a model
  - Send the monitoring parameter to prometheus gate
  - Create visualization on the Grafana
  - Visualize the details on the Screen.

prasannavj@gmail.com

LIVE DEMO – SOURCE CODE

- Reference from demo github link
  - (Source Credit) https://github.com/sohiniroych/ML-Monitoring-with-Grafana
  - Source code and steps for configuration and management.
  - https://github.com/prasannavj/MACHINELEARNINGPROBLEMS