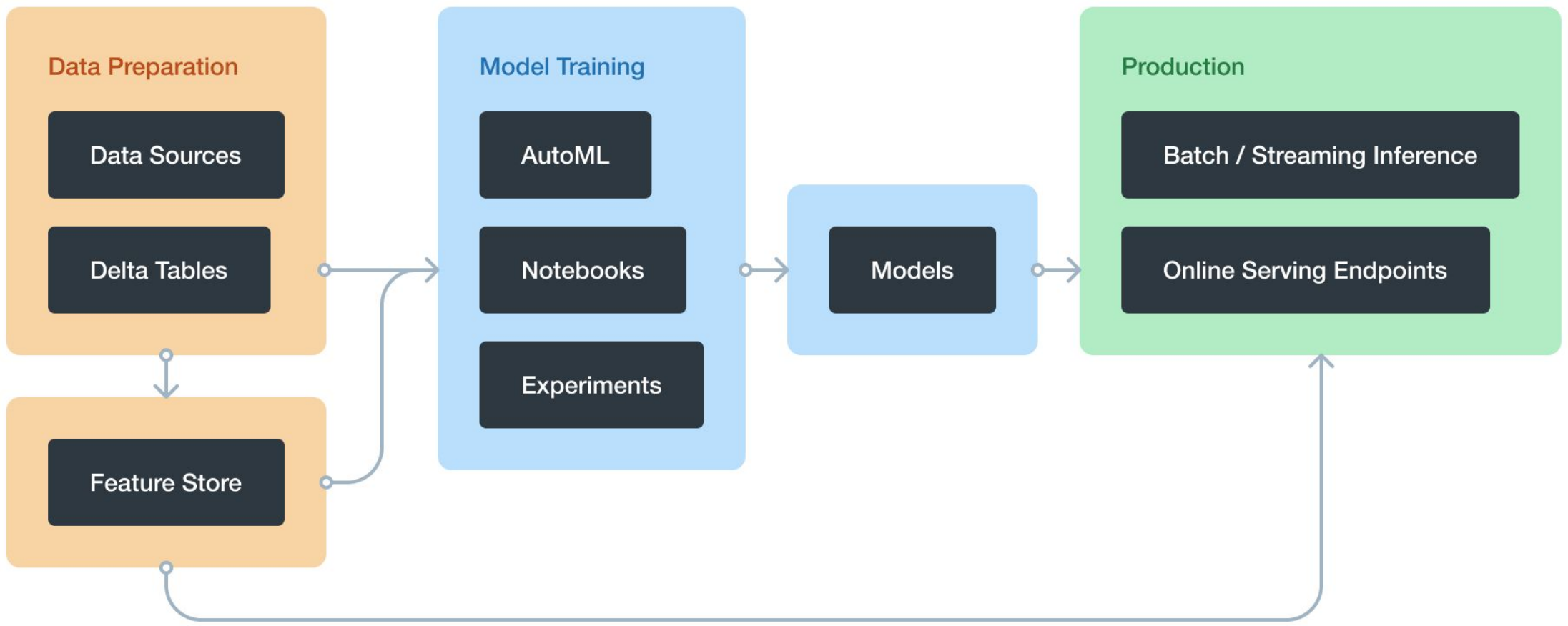
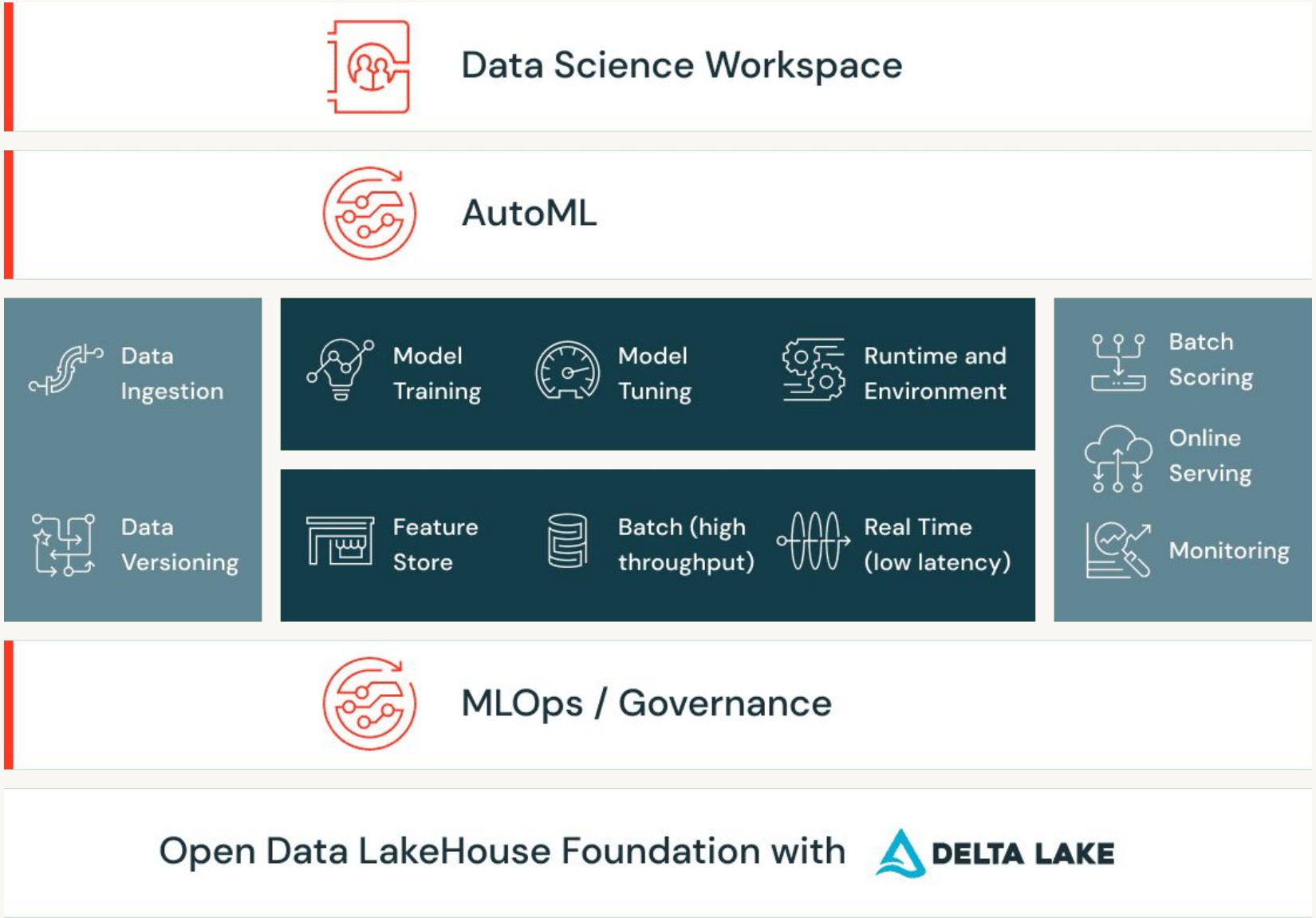
databricks

Model Training with Databricks ML

Learning objectives

- Describe how Databricks ML supports manual and automatic model development.





Hold for Video

Required parameters:

- `dataset` - Input Spark or pandas DataFrame that contains training features and targets. If using a Spark DataFrame, it will convert it to a Pandas DataFrame under the hood by calling .toPandas() - just be careful you don't OOM!
- `target_col` - Column name of the target labels

We will also specify these optional parameters:

- `primary_metric` - Primary metric to select the best model. Each trial will compute several metrics, but this one determines which model is selected from all the trials. One of `r2` (default, R squared), `mse` (mean squared error), `rmse` (root mean squared error), `mae` (mean absolute error) for regression problems.
- `timeout_minutes` - The maximum time to wait for the AutoML trials to complete. `timeout_minutes=None` will run the trials without any timeout restrictions
- `max_trials` - The maximum number of trials to run. When `max_trials=None`, maximum number of trials will run to completion.

```
from databricks import automl

summary = automl.regress(train_df, target_col="price", primary_metric="rmse", timeout_minutes=5, max_trials=10)
```

databricks

Model Training with Databricks ML

Use AutoML to develop a baseline model

Learning objectives

- Use AutoML to develop a baseline model.