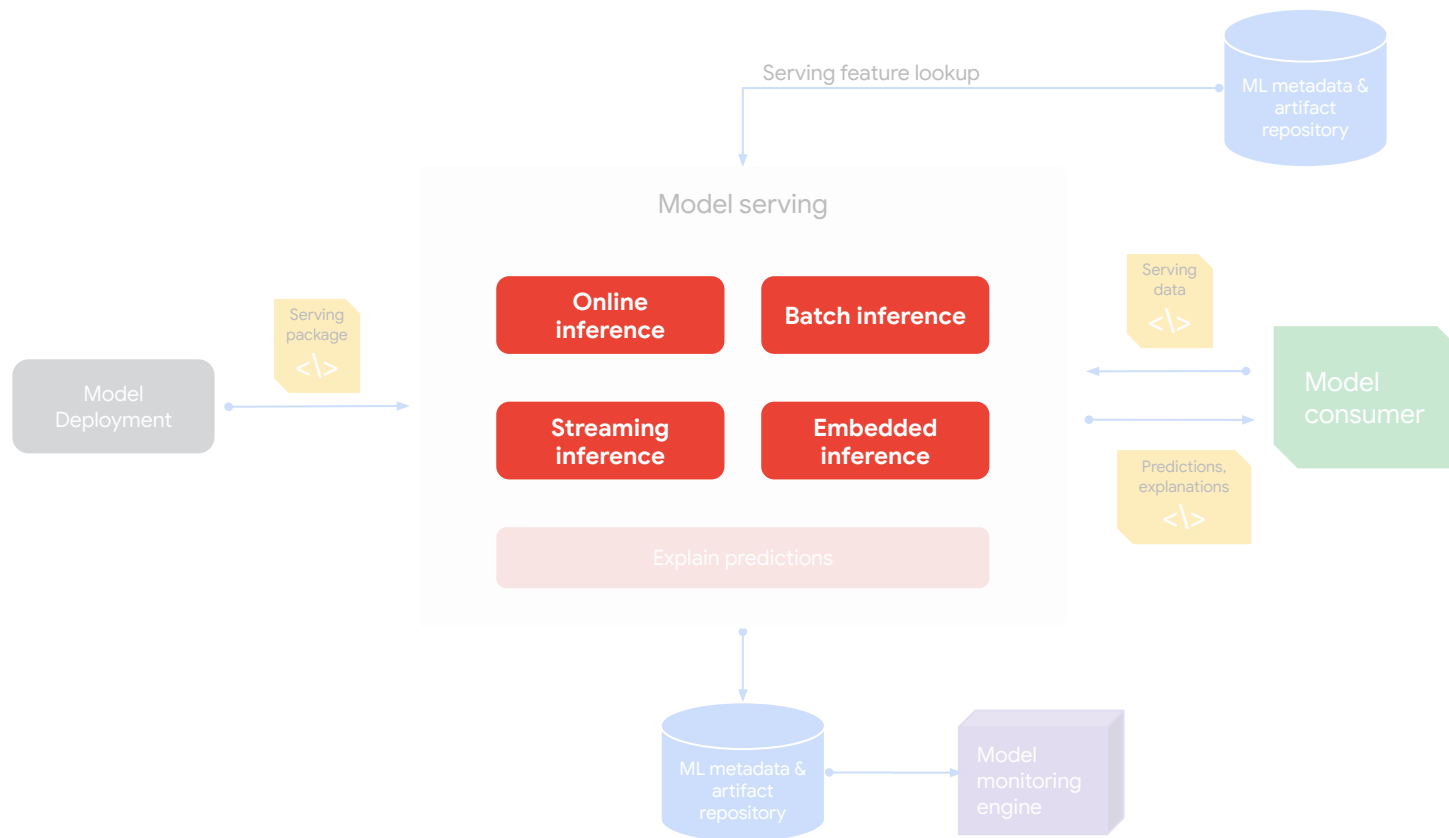


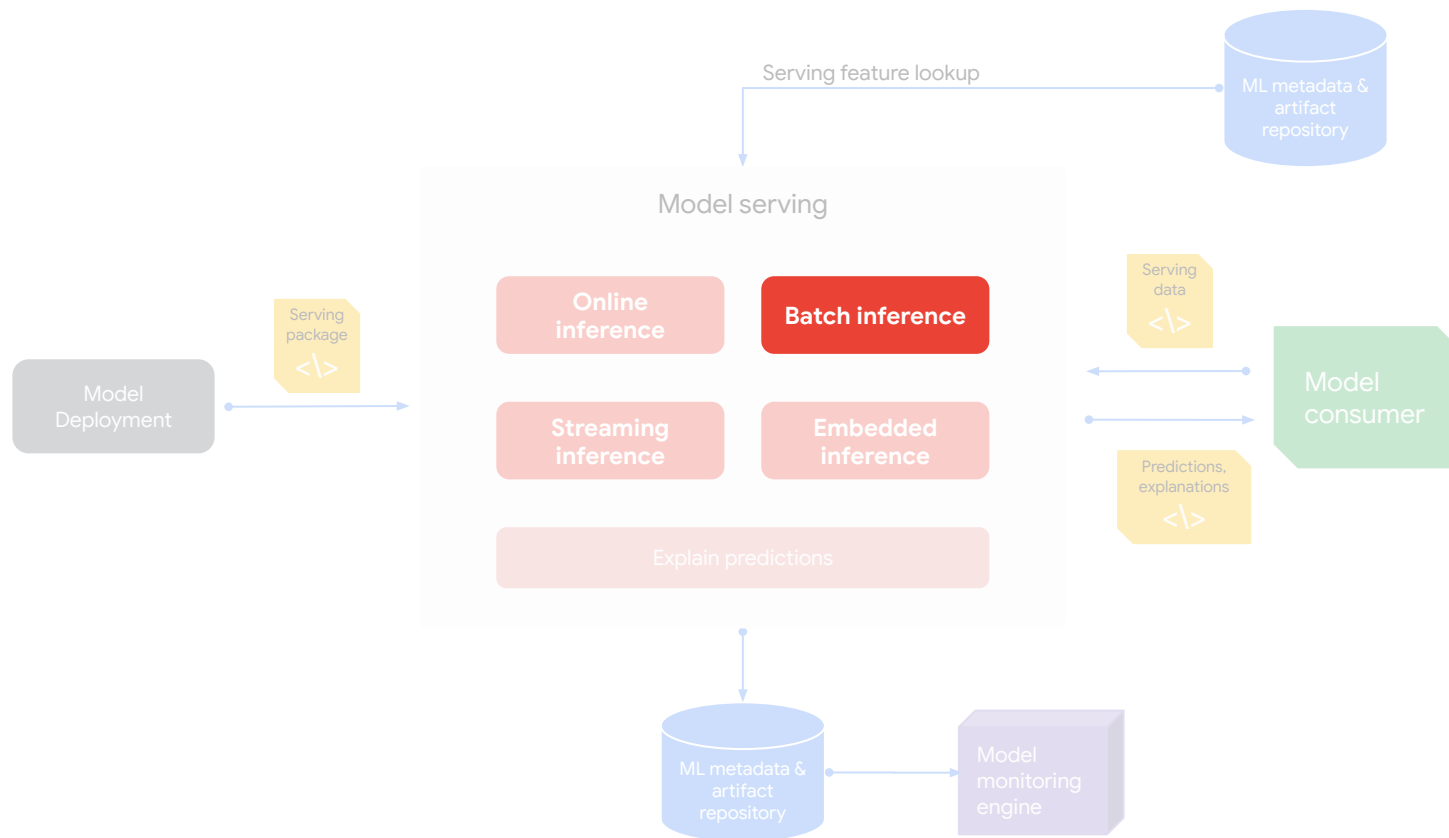
Prediction Serving Scenarios: Batch



MLOps: Prediction Serving



MLOps: Prediction Serving



The MLOps Personas



ML
Engineer



ML
Researcher



Data
Scientist



Data
Engineer



Software
Engineer



DevOps



Business
Analyst

Batch Inference:

What is it?

- Batch inference is the process of generating predictions on a batch of observations.



Batch Inference:

What is it?

- Batch inference is the process of generating predictions on a batch of observations.
- The batch jobs are typically generated on some recurring schedule (e.g. hourly, daily)



Batch Inference:

What is it?

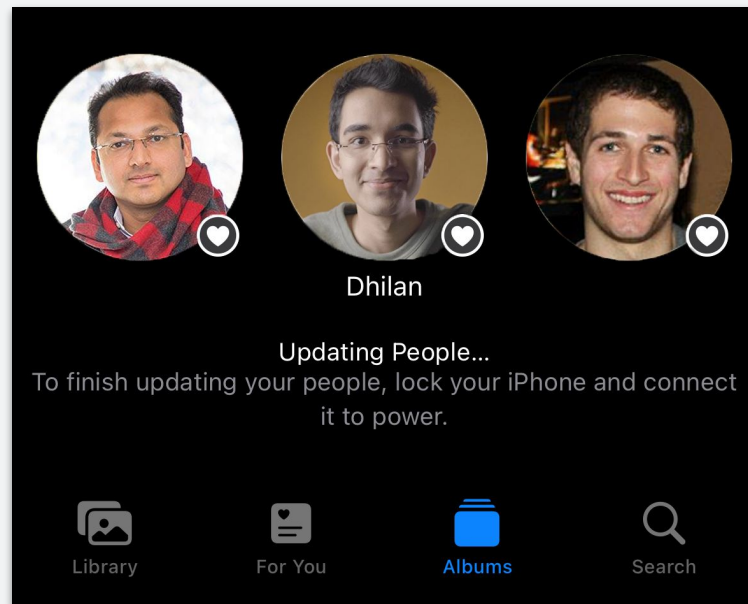
- Batch inference is the process of generating predictions on a batch of observations.
- The batch jobs are typically generated on some recurring schedule (e.g. hourly, daily)
- It is easily scalable across a large number of systems



Batch Inference:

When is it useful?










- Face recognition on your Photos app



Batch Inference:

When is it useful?

- Face recognition on your Photos app
- Recommendations on movies to watch based on past shows

Batch Inference:

How it works?

The batch jobs are typically generated on some recurring schedule (e.g. hourly, daily)

Observations

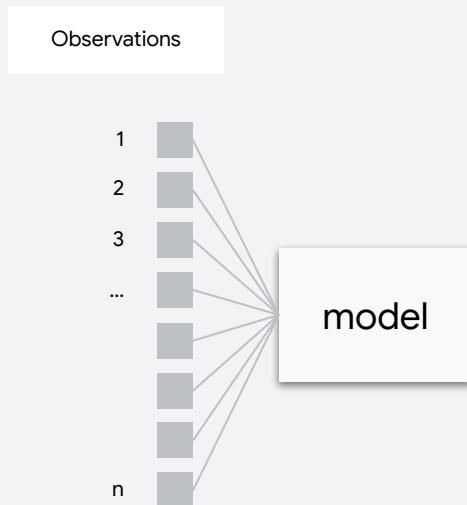
1
2
3
...
n

A vertical column of gray squares, each representing an observation in a batch. The squares are aligned to the right of the labels 1, 2, 3, ..., n.

Batch Inference:

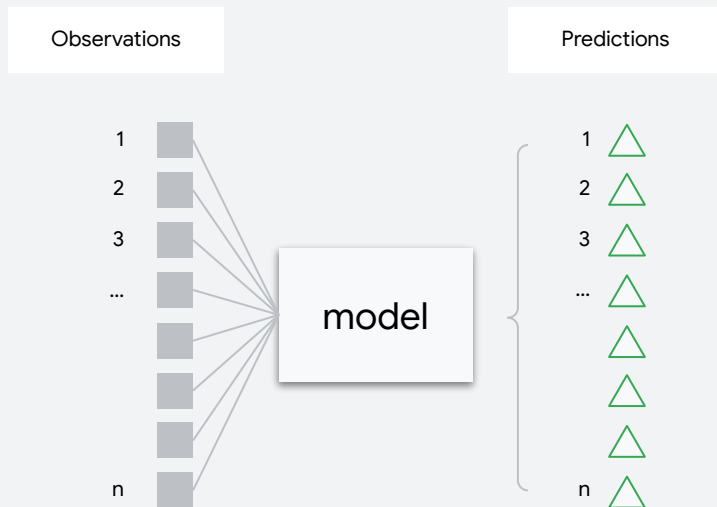
How it works?

The batch jobs are typically generated on some recurring schedule (e.g. hourly, daily)



Batch Inference: How it works?

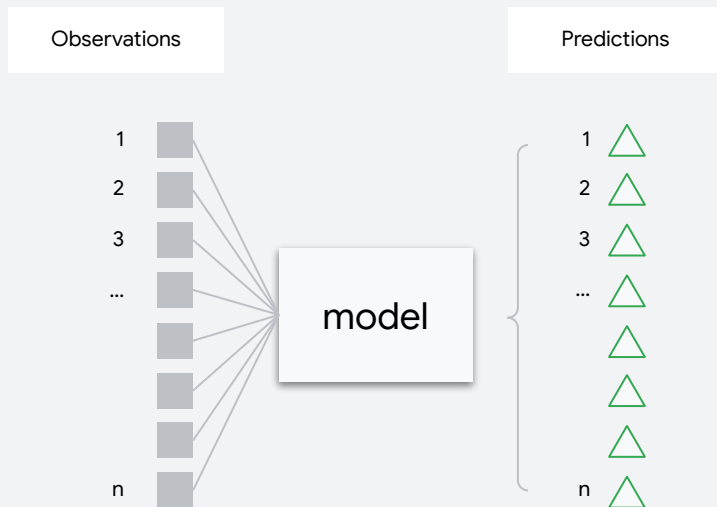
The batch jobs are typically generated on some recurring schedule (e.g. hourly, daily)



Batch Inference: How it works?

The batch jobs are typically generated on some recurring schedule (e.g. hourly, daily)

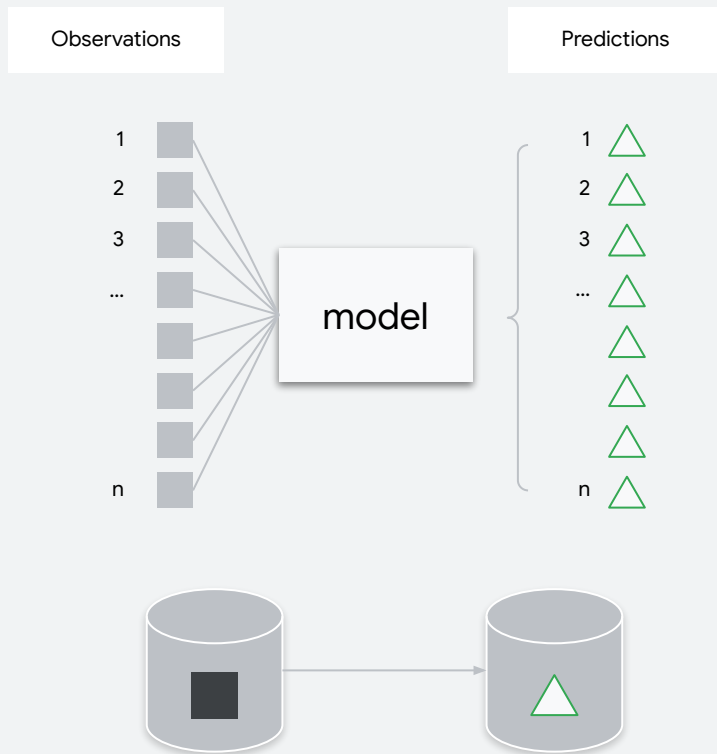
These predictions can be stored in a database and can be made available to developers or users



Batch Inference: How it works?

The batch jobs are typically generated on some recurring schedule (e.g. hourly, daily)

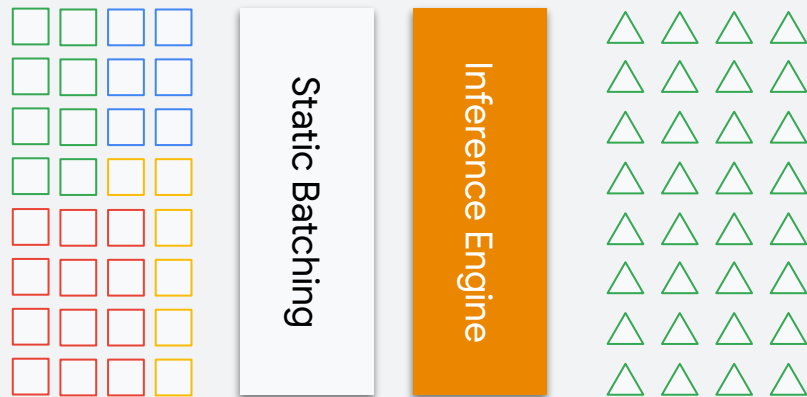
These predictions can be stored in a database and can be made available to developers or users



Batch Inference:

What metrics?

- No real-time deadlines
- Throughput mode
- Queries



Batch Inference:

Pros & Cons

Pros

- + Don't need to worry much about **latency cost**

Batch Inference:

Pros & Cons

Pros

- + Don't need to worry much about **latency cost**
- + Can likely use **batch quota** or some giant MapReduce

Batch Inference:

Pros & Cons

Pros

- + Don't need to worry much about **latency cost**
- + Can likely use **batch quota** or some giant MapReduce
- + Can do **post-verification** of predictions before pushing

Batch Inference:

Pros & Cons

Pros

- + Don't need to worry much about **latency cost**
- + Can likely use **batch quota** or some giant MapReduce
- + Can do **post-verification** of predictions before pushing

Cons

- **Cold start** penalty is unavoidable

Batch Inference:

Pros & Cons

Pros

- + Don't need to worry much about **latency cost**
- + Can likely use **batch quota** or some giant MapReduce
- + Can do **post-verification** of predictions before pushing

Cons

- **Cold start** penalty is unavoidable
- Can only predict things we know about — **bad for long tail**

Batch Inference:

Pros & Cons

Pros

- + Don't need to worry much about **latency cost**
- + Can likely use **batch quota** or some giant MapReduce
- + Can do **post-verification** of predictions before pushing

Cons

- **Cold start** penalty is unavoidable
- Can only predict things we know about — **bad for long tail**
- Update **latency** is likely measured in hours or days

Batch Inference:

Pros & Cons

Pros

- + Don't need to worry much about **latency cost**
- + Can likely use **batch quota** or some giant MapReduce
- + Can do **post-verification** of predictions before pushing

Cons

- **Cold start** penalty is unavoidable
- Can only predict things we know about — **bad for long tail**
- Update **latency** is likely measured in hours or days

Scenario

Metric



Batch inference
(e.g. photo sorting app)

Throughput