BITS Pilani
Pilani | Dubai | Goa | Hyderabad

Batch Prediction Versus Online Prediction

Pravin Y Pawar

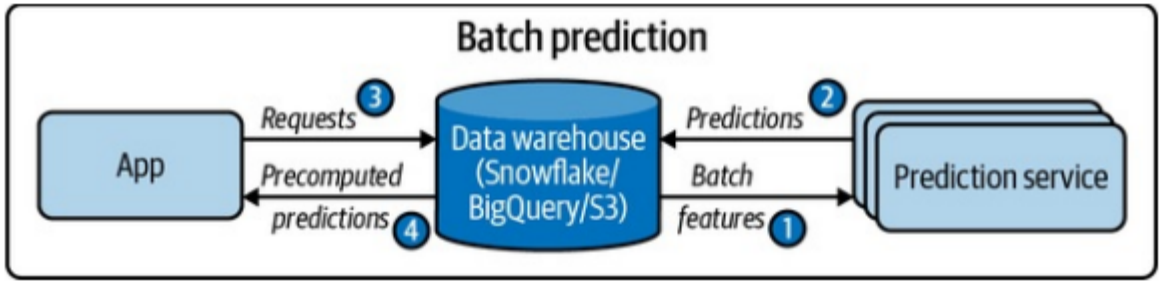
Adapted from ‘Designing Machine Learning Systems”

Three main modes of prediction

- Batch prediction, which uses only batch features
- Online prediction that uses only batch features (e.g., precomputed embeddings)
  - Aka on-demand prediction
- Online prediction that uses both batch features and streaming features
  - aka streaming prediction

Batch prediction

- When predictions are generated periodically or whenever triggered
  - Predictions are stored somewhere, such as in SQL tables or an in-memory database, and retrieved as needed
- For example,
  - Netflix might generate movie recommendations for all of its users every four hours,
  - the precomputed recommendations are fetched and shown to users when they log on to Netflix
- AKA asynchronous prediction: predictions are generated asynchronously with requests



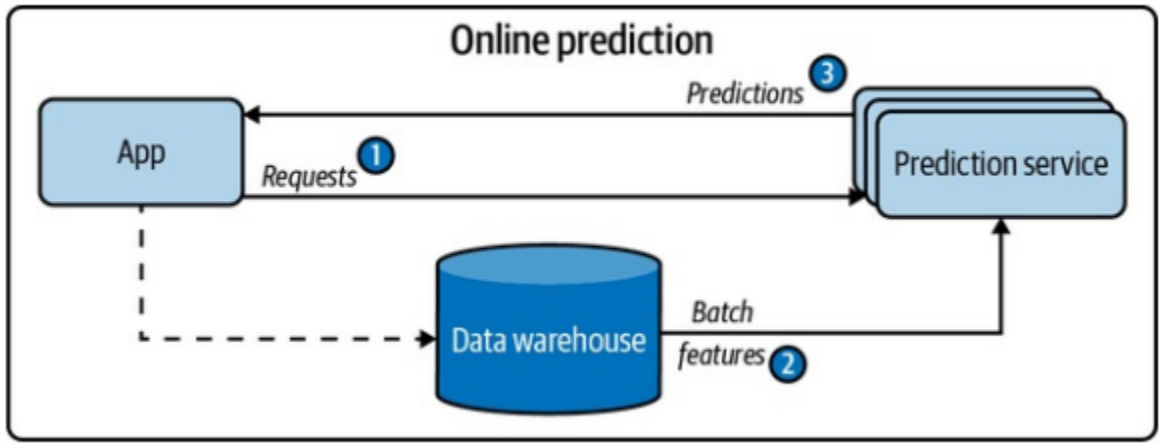
A simplified architecture for batch prediction

Online prediction

Using only batch features

- When predictions are generated and returned as soon as requests for these predictions arrive
  - For example, enter an English sentence into Google Translate and get back its French translation immediately
  - aka on-demand prediction

- Traditionally, when doing online prediction, requests are sent to the prediction service via RESTful APIs
  - aka synchronous prediction: predictions are generated in synchronization with requests



A simplified architecture for online prediction that uses only batch features

Online vs Batch prediction

Some key differences between batch prediction and online prediction

	Batch prediction (asynchronous)	Online prediction (synchronous)
Frequency	Periodical, such as every four hours	As soon as requests come
Useful for	Processing accumulated data when you don't need immediate results (such as recommender systems)	When predictions are needed as soon as a data sample is generated (such as fraud detection)
Optimized for	High throughput	Low latency

Online prediction

Using both batch and streaming features - “streaming prediction”

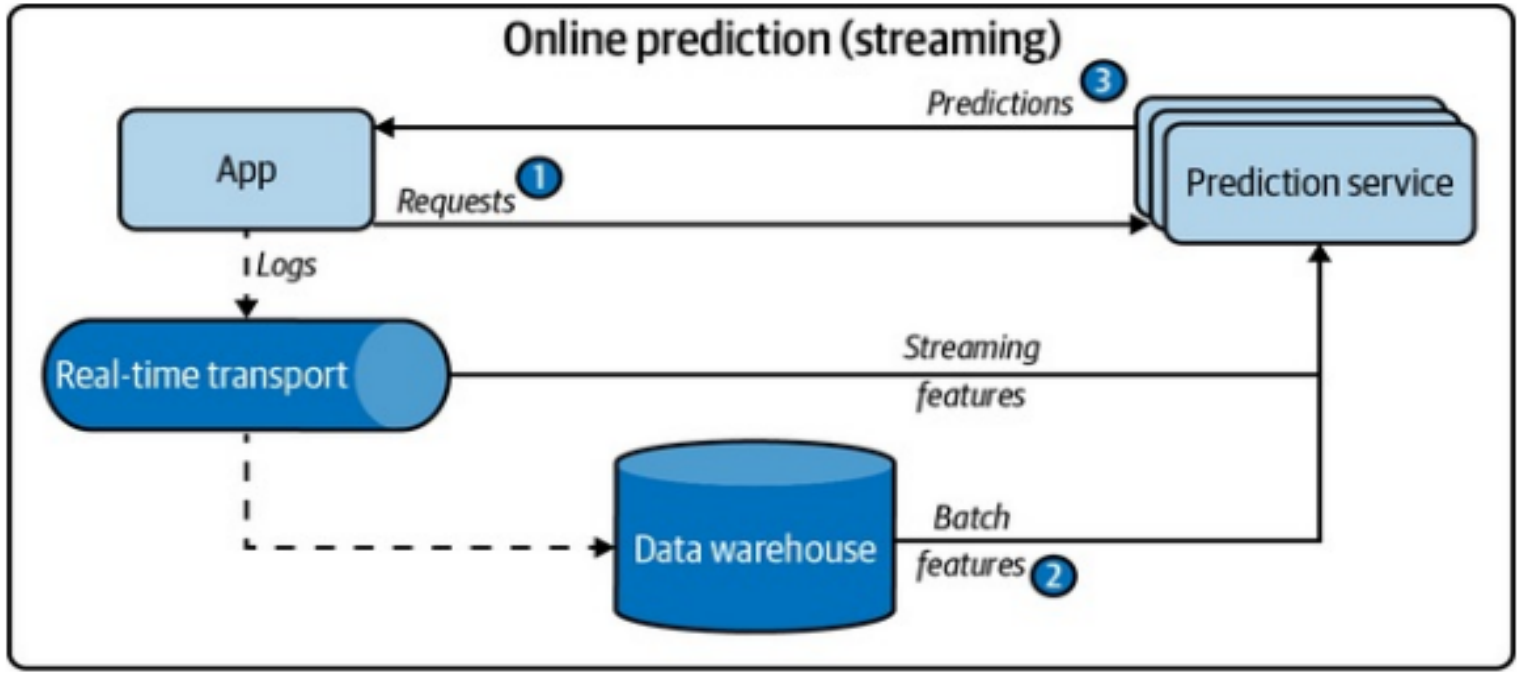
- Batch features
  - features computed from historical data, such as data in databases and data warehouses
  - for e.g. a restaurants rating

- Streaming features
  - Features computed from streaming data — data in real-time transports
  - for e.g. estimation for delivery time

- In batch prediction, only batch features are used
- In online prediction, however, it’s possible to use both batch features and streaming features.
  - For example, after a user puts in an order on DoorDash, they might need the following features to estimate the delivery time:
    - The mean preparation time of this restaurant in the past
    - In the last 10 minutes, how many other orders they have, and how many delivery people are available

Streaming prediction

A simplified architecture for online prediction that uses both streaming features and batch features



A simplified architecture for online prediction that uses both batch features and streaming features

Thank You!

In our next session: