Streaming data into your Storage Cluster.

Kafka – Publish/subscribe

How does new data get into your cluster? Especially if it’s ’Big data’?

* New log entries from your webservers
* New sensor data from your IoT system
* New stock trades

Streaming lets you publish this data, in real time, to your cluster.

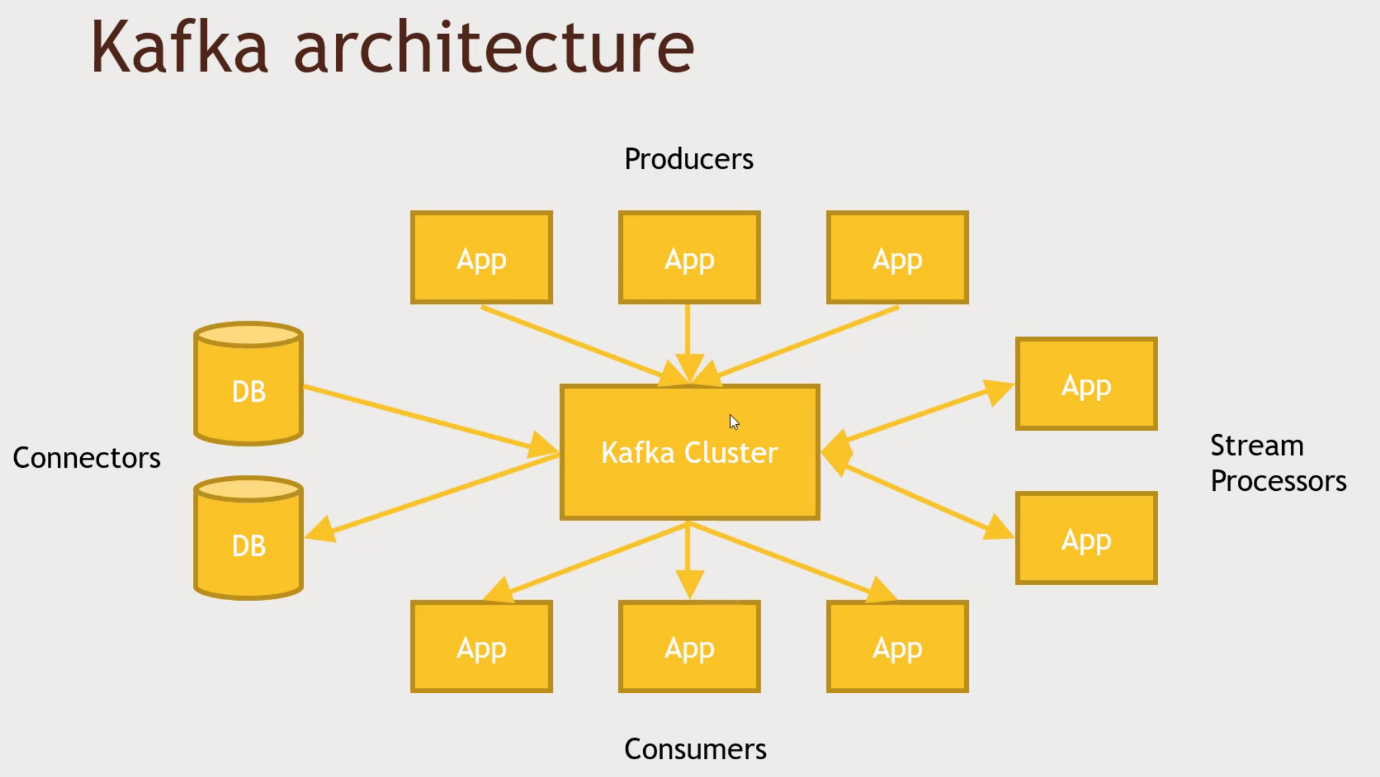
* And you can even process it in real time as it comes in!

Two problems

1. How to get data from many different sources flowing into your cluster
2. Processing it when it gets there

HOW TO GET DAT FROM MANY SOURCES

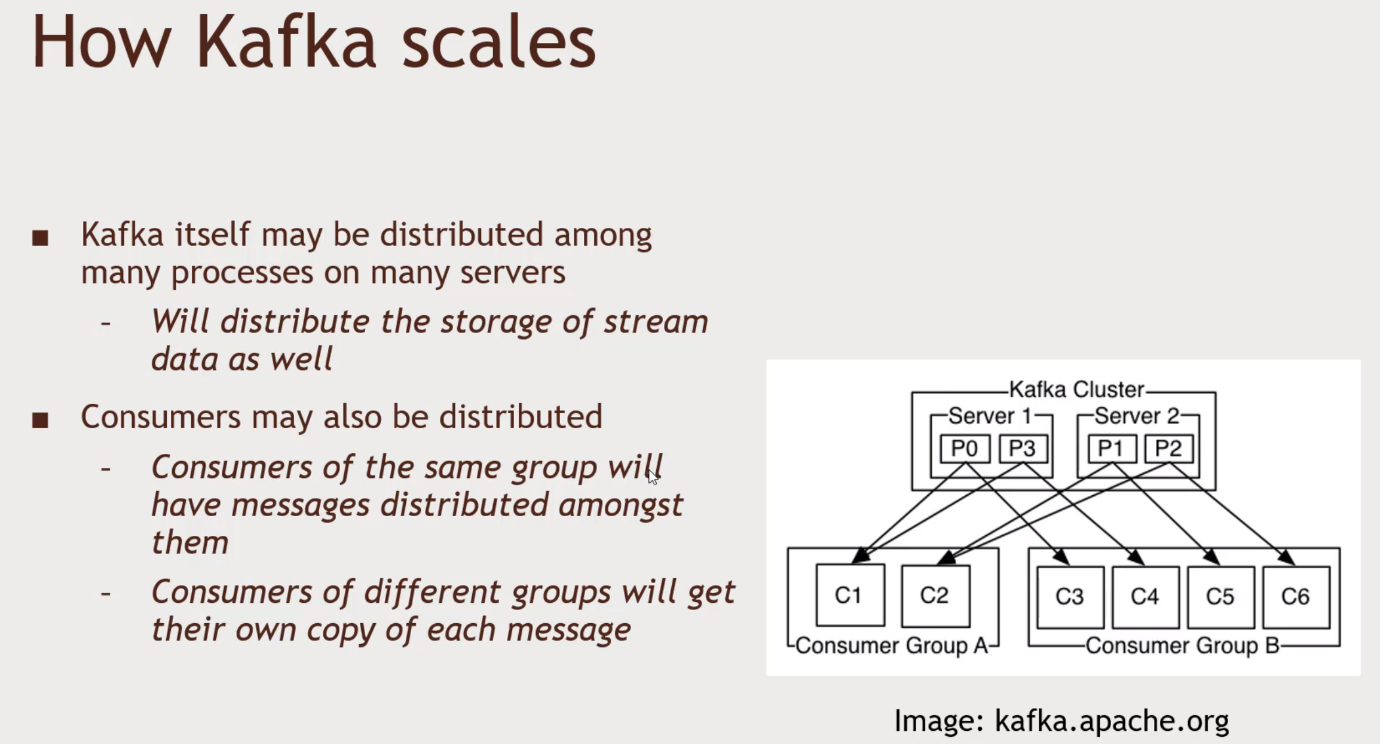
Enter Kafka

* General-purpose **publish/subscribe messaging system**
* Kafka servers **store all incoming messages from publishers** for some period of time, and publishes them to a stream of data called a **topic. Specific stream weblogs, sensor data.**
* Kafka consumers subscribe to one or more topics, and receive data as it’s published
* A stream/topic can have many different consumers, all with their own position in the stream maintained
* It’s not just for Hadoop

Producers – Generating data, individual apps, log lines, sensor data. Pushing data into your Kafka Customer

Consumers – Receiving the data. Read the data and process it in some way. Spark Streaming app, specific topic that receive data from the producers. You can use off the shelf stuffs

Stream processors- Transform data as it comes in. Unstructured weblines, that listens for new log lines and extract info.

Connectors – Plugin modules for various databases, that publish new rows in some new DataBase tables. Or can receive new topics from Kafka, Publish or Receive.