Cluster Redundancy and Real-Time Data

Redundancy vs. Failure Recovery

Redundancy

- Data exists multiple places

Failure Recovery

- Data can be retrieved if lost or destroyed

Redundancy Operations

- Election of a Master Node
- Detection of crashes
- Communication about failures
- Determination of what’s available when
- Creation of metadata to track

Zookeeper

- Creates redundancy for the master node

- Help you recover from:
  - Hard drive failure
  - Loss of power
  - Drift: computers out of sync
  - Time zone issues

znodes

Persistent

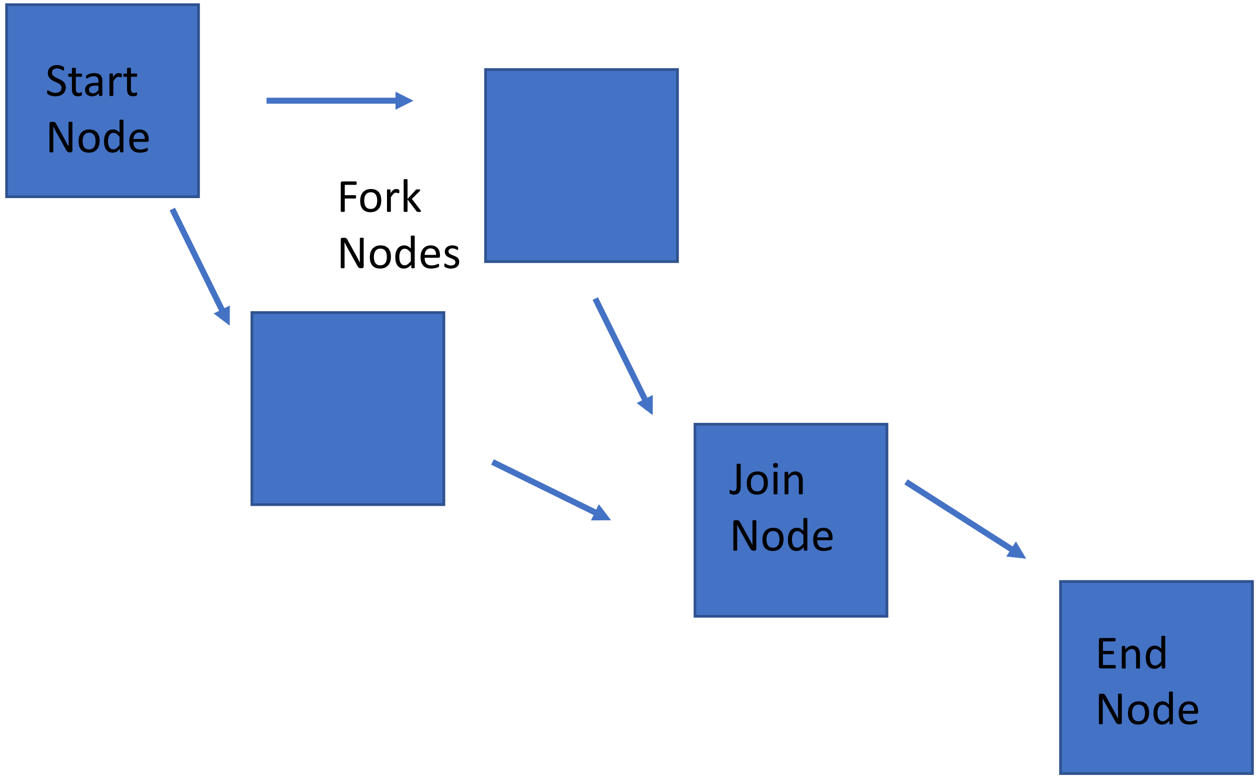
- Always around just in case

Ephemeral

- Only created when you have an issue

Oozie

- Cluster management software
- Chain operations together



Streaming Data

- Accessing and using data in real time

- Data dumps straight from the generation point

IoT – Internet of Things

- “Smart” technology that connect to the web

- Generates machine data

Streaming Software

- Kafka: data stored until you pick it up
- Flume: data flows to your end destination
- Spark Streaming: data arrives in microbatches
- Storm: real-time processing
- Flink: faster real-time processing & uses an API

Windows & Intervals

- Window – snapshot of streamed data
- Batch interval – how often data comes in
- Slide interval – how often you use data in a window
- Window interval – how far back in time you get data

Questions?