A diagram of a system architecture

Description automatically generated

Docker-compose.yml

1. **Version**: This file uses Docker Compose version 3 for service orchestration.
2. **Common Spark Configuration**: The x-spark-common section defines a reusable Spark worker configuration using the bitnami/spark:latest image, specifying worker settings such as cores, memory, and master URL.
3. **Zookeeper Service**:
   * Uses confluentinc/cp-zookeeper:7.4.0 image.
   * Exposes port 2181 for client connections.
   * Configured with a health check to ensure Zookeeper is running properly.
4. **Kafka Broker Service**:
   * Uses confluentinc/cp-server:7.4.0 image.
   * Depends on the Zookeeper service for coordination.
   * Exposes ports 9092 (client communication) and 9101 (JMX metrics).
   * Configured to advertise listeners for internal and external connections.
5. **Spark Master Service**:
   * Uses bitnami/spark:latest image.
   * Runs the Spark master process (org.apache.spark.deploy.master.Master).
   * Exposes ports 9090 for the Spark Web UI and 7070 for the Spark master communication.
6. **Spark Worker Services**:
   * spark-worker-1 and spark-worker-2 inherit the configuration defined in x-spark-common.
   * These workers connect to the Spark master and are part of the same network.
7. **Network Configuration**:
   * All services are connected to the datamasterylab network, allowing them to communicate with each other.
   * The network is implicitly created, and its driver is the default Docker bridge.
8. **Health Checks**:
   * Both Zookeeper and Kafka broker services have health checks to monitor their status and ensure they are available before starting dependent services.
9. **Volumes**:
   * Spark services mount a ./jobs directory on the host to the /opt/bitnami/spark/jobs directory in the container, enabling access to Spark jobs.
10. **Port Mapping**:
    * Specific ports are mapped between the host and container, allowing access to Spark Web UI, Kafka, and Zookeeper from the host machine.



Now install pip install confluent\_kafka simplejson pyspark