**CS 4287/5287: Principles of Cloud Computing**
**Programming Assignment #4**
Handed out: 11/07/2024; Due: Sunday, 12/08/2024 (11:59 pm CDT on Brightspace)
**Theme**: MapReduce using Apache Spark-based Batch Processing

**Team 14 Members**:
Maddox
Emma
Abhay

---

## Project Components and Team Responsibilities

### Spark Batch Processing for Inference Counting

**Apache Spark Setup (Completed by: Maddox)**
**Tasks Completed**:

- Installed Apache Spark on the VM cluster, configured environment variables, and verified Spark functionality.
- Implemented MapReduce logic using PySpark to count incorrect inferences on a per-producer basis.
- Integrated Spark job with the existing database to fetch and process stored inference data.
  **Deliverables**:
- Documented Spark setup and PySpark script for batch processing.
- Verified correctness of MapReduce results with database entries.

### Data Collection and Kafka Integration

**Data Ingestion (Completed by: Abhay)**
**Tasks Completed**:

- Scaled data ingestion by deploying multiple producers to publish large volumes of images to Kafka.
- Collected substantial batches of data to be used in the Spark MapReduce application.
  **Deliverables**:
- Documented producer scaling and data collection methodology.
- Validated database entries for use in batch processing.

**Kubernetes Scaling on CH-819381**

**Cluster Scaling and Deployment (Completed by: Maddox)**
**Tasks Completed**:

- Migrated the solution to pre-allocated Kubernetes clusters on CH-819381 after validation on the 4-VM setup.
- Deployed pods in the team's namespace, leveraging private Docker registries on the cluster masters.
- Configured firewall rules and network policies for cross-cluster communication.
  **Deliverables**:
- Documented scaling steps and configuration changes for CH-819381.
- Verified pod deployments and data flow across clusters.

## General Collaboration and Integration

**Documentation and Reporting (Completed by: Emma)**:
Compiled a comprehensive report detailing project setup, testing processes, and MapReduce analysis results.

**Video Demonstration (Completed by: Maddox)**:
Created a video demonstration showcasing the batch and streaming processing components.

---

## Final Deliverables

1. **Code**: PySpark scripts, Kubernetes YAML files, Dockerfiles
2. **Documentation**: Setup steps, testing reports
3. **Video**: Detailed demonstration of project components and functionality.