제2장 Spark의 설계와 운행 원리

2.4 Spark의 배포와 응용 방식

강사: 이양원 교수
연락처: ywlee1212@gmail.com

Slide 1: 강의 소개

안녕하세요, 오늘은 제2장 Spark의 설계와 운행 원리에 대해 학습하겠습니다. 저는 이양원 교수입니다. 궁금한 점이 있으시면 이메일이나 홈페이지를 통해 연락해주세요.

Slide 1:

大家好，今天我们来学习第2章 Spark的设计与运行原理。我是李阳源老师，如果大家有任何问题，可以通过邮箱 ywlee1212@gmail.com 联系我，或者访问我的主页。

Slide 2: 제2장 학습 시작

이제 제2장 학습을 정식으로 시작하겠습니다. 이 장의 주제는 "Spark의 설계와 운행 원리"입니다. 이 장에서는 Spark라는 빅데이터 처리 프레임워크의 핵심 개념과 작동 메커니즘을 깊이 있게 이해해보겠습니다.

Slide 2:

现在我们正式开始第2章的学习，这一章的主题是"Spark的设计与运行原理"。在这一章中，我们将深入了解Spark这个大数据处理框架的核心概念和工作机制。

Slide 3: 2.4절 - Spark의 배포와 응용 방식

다음으로 학습할 내용은 2.4절: Spark의 배포와 응용 방식입니다.

이는 Spark의 실제 응용을 이해하는 중요한 내용입니다.

Slide 3: 2.4 Spark의 부서와 응용 방식

接下来我们要学习的是2.4节：Spark的部署和应用方式。

这是理解Spark实际应用的重要内容。

Slide 4: 단일 노드와 클러스터

Spark는 두 가지 다른 배포 방식을 가질 수 있습니다:

- **단일 노드 배포**: 개발 테스트 환경에 적합
- **클러스터 배포**: 대규모 데이터를 처리하는 프로덕션 환경에 사용

Slide 5: 클러스터 배포의 세 가지 방식

클러스터 배포의 경우, Spark는 세 가지 주요 클러스터 관리 방식을 제공합니다:

1. **Standalone**
2. **YARN**
3. **Mesos**

각 방식마다 고유한 특징과 적용 시나리오가 있습니다.

Slide 6: Standalone 방식

먼저 Standalone 방식을 살펴보겠습니다.

- Spark에 내장된 클러스터 관리자
- 설정이 비교적 간단
- 효율성이 높지 않음

적합한 용도: 소규모 클러스터나 학습 테스트

Slide 4: - 단기와 집군

Spark可以有两种不同的部署方式：

单机部署和集群部署:

- 单机部署适合开发测试环境，
- 而集群部署则用于生产环境中处理大规模数据。

Slide 5: 5 - 집군 배포의 세 가지 방식

对于集群部署，Spark提供了三种主要的集群管理方式：

1. Standalone、
2. YARN
3. 和Mesos。

每种方式都有各自的特点和适用场景。

Slide 6: - Standalone 방식

首先我们来看Standalone方式。

- 这是Spark自带的集群管理器，
- 配置相对简单，
- 但是效率不高。

它适合小规模集群或者学习测试使用。

Slide 7: Mesos 방식

다음은 Mesos 방식입니다.

- Spark와 Mesos 모두 아파치 재단의 프로젝트
- 혈연관계가 있어 성능 매칭이 매우 좋음
- 더 나은 리소스 관리와 태스크 스케줄링 능력 제공

Slide 7: - Mesos 方式

接下来是Mesos方式。

- 由于Spark和Mesos都是Apache基金会的项目，
- 它们之间有血缘关系，因此性能匹配度（pǐ pèi dù 호환성수준）很好。
- Mesos提供了更好的资源管理和任务调度能力。

Slide 8: YARN 방식 소개

세 번째 방식은 YARN입니다.

- Hadoop 생태계의 중요한 구성 요소
- 다양한 컴퓨팅 프레임워크와 결합 가능:
    - Spark
    - MapReduce
    - Storm 등

Slide 8: 8 - YARN 方式 소개

第三种方式是YARN。

- YARN是Hadoop生态系统的重要组成部分，
- 可以与Spark、MapReduce、Storm等多种计算框架结合使用。

Slide 9: YARN 생태계 구조

이 그림에서 볼 수 있듯이,

YARN 관리 하에서:

- **하위 계층**: HDFS 분산 파일 시스템
- **중간 계층**: YARN 리소스 관리 계층
- **상위 계층**: Spark 코어 엔진
- **최상위 계층**: Spark의 각 컴포넌트들
    - **Spark SQL**: 쿼리 계산용
    - **Spark Streaming**: 스트림 계산용
    - **MLlib**: 머신러닝용
    - **GraphX**: 그래프 계산용

Slide 9: - YARN 생태계 구조

从这个图中我们可以看到，

在YARN管理下，

- 底层是HDFS分布式文件系统，
- 中间是YARN资源管理层，
- 上层是Spark核心引擎，
- 最上层是Spark的各个组件：

1. Spark SQL用于查询计算，
2. Spark Streaming用于流计算，
3. MLlib用于机器学习，
4. GraphX用于图计算。

Slide 10: 2015년 이후의 변화

2015년부터 Spark는 점차 빅데이터 처리의 주류 선택이 되었습니다.

- 많은 시나리오에서 Hadoop MapReduce를 대체하기 시작
- 기업 빅데이터 처리의 첫 번째 선택이 됨

Slide 10: - 2015년 이후의 변화

从2015年开始，Spark逐渐(**zhú jiàn**)成为大数据处理的主流选择，

- 在很多场景下开始替代Hadoop MapReduce，
- 成为企业大数据处理的首选框架。

Slide 11: Spark의 우세

Spark의 우세는 주로 두 가지 측면에서 나타납니다:

1. **낮은 지연시간**: 처리 속도가 빠름
2. **빠른 배치 처리**: 대량의 데이터를 신속하게 처리 가능

이러한 우세로 인해 Spark는 빅데이터 처리 분야에서 강한 경쟁력을 갖고 있습니다.

Slide 12: Spark vs MapReduce

구체적으로 말하면, Spark는 배치 처리에서 MapReduce보다 더 나은 성능을 얻습니다.

이유:

- Spark의 메모리 컴퓨팅 특성
- 더 최적화된 실행 엔진

Slide 11: - Spark的优势

Spark的优势主要体现在两个方面：

1. 第一是延迟低，处理速度快；
2. 第二是批处理快(**pī chǔ lǐ)**，能够快速处理大量数据。

这些优势使得Spark在大数据处理领域具有很强的竞争力。

Slide 12: - Spark vs MapReduce

具体来说，Spark做批处理比MapReduce获得更好的性能。

这主要得益于Spark的内存计算特性和更优化的执行引擎。

Slide 13: Spark의 다양한 컴포넌트

Spark의 또 다른 우세는 컴포넌트의 다양화입니다.

SparkContext를 핵심으로:

- **Spark SQL**: 쿼리 계산 담당
- **Spark Streaming**: 스트림 계산 담당
- **MLlib**: 머신러닝 기능 제공
- **GraphX**: 그래프 계산 능력 제공

이러한 통합된 아키텍처로 개발자들은 하나의 플랫폼에서 다양한 유형의 데이터 처리 작업을 완료할 수 있습니다.

Slide 13:- Spark组件多样化组件多样化

Spark的另一个优势是组件多样化。

以SparkContext为核心，

- Spark SQL负责查询计算，
- Spark Streaming负责流计算，
- MLlib提供机器学习功能，
- GraphX提供图计算能力。

这种统一的架构让开发者可以在一个平台上完成多种类型的数据处理任务。

Slide 14: Spark 2.3과 Structured Streaming

Spark 2.3 버전부터 Structured Streaming이 도입되었습니다.

특징:

- 지속적 스트림 모드
- 밀리초 단위의 스트림 계산 실현
- Spark의 스트림 처리 분야에서의 경쟁력을 크게 향상
- Apache Flink 등 전문 스트림 처리 프레임워크와 경쟁 가능

Slide 14: - Spark 2.3과 Structured Streaming

从Spark 2.3版本开始，

引入了Structured Streaming，

- 这是一种持续流模式(chíxù liú mó shì)，
- 能够实现毫秒级(**háo miǎo jí)**流计算，
- 大大提升(**tí sheng)**了Spark在流处理方面的竞争能力，
- 可以与Apache Flink等专业流处理框架竞争**(jìng zhēng)**。

Slide 15: Hadoop 생태계와의 관계

Spark가 여러 방면에서 MapReduce보다 우수하지만, 이것이 Spark가 Hadoop 생태계를 완전히 대체한다는 것을 의미하지는 않습니다.

실제로:

- Spark는 종종 Hadoop의 다른 컴포넌트들과 협력하여 작업
    - HDFS
    - HBase 등

Slide 15: - Hadoop 생태계와의 관계

虽然Spark在很多方面优于MapReduce，

但这并不意味着Spark要完全替代Hadoop生态系统。

实际上，

- Spark经常与Hadoop的其他组件如HDFS、HBase等协同(**xié tóng)**工作。

Slide 16: 통합 생태계

최종적으로 Spark, Hadoop, HDFS, HBase 등의 기술들은 통합된 빅데이터 생태계를 형성합니다.

- 이들이 함께 기업의 관련 응용 시나리오 요구를 만족
- 각 컴포넌트마다 특정한 기능과 우세가 있음
- 합리적인 조합을 통해 강력하고 완전한 빅데이터 솔루션 구축 가능

Slide 16: - 통합 생태계

最终，Spark、Hadoop、HDFS、HBase等技术会形成一个统一的大数据生态系统，

它们共同来满足企业的相关应用场景需求。

每个组件都有其特定的功能和优势，

通过合理的组合可以构建出强大而完整的大数据解决方案。

연습 문제

Slide 17 **문제 1**

Spark 클러스터 배포 모드 중에서 Spark에 내장된 클러스터 관리자는 어느 것입니까?

정답: C (Standalone)

Slide 18**문제 2**

다음 중 Spark의 핵심 컴포넌트가 아닌 것은?

정답: D

Slide 19: **문제 3**

Spark와 Mesos 간의 성능 매칭이 좋은 주요 이유는 무엇입니까?

정답: B (모두 아파치 재단 프로젝트 / 혈연관계)

Slide 20:**문제 4**

Spark 2.3 버전에서 도입된 새로운 기능은 무엇입니까?

정답: C (Structured Streaming)

Slide 21: **문제 5**

YARN이 관리하는 Spark 생태계에서 하위 계층에 사용되는 분산 파일 시스템은 무엇입니까?

정답: A (HDFS)

연습 문제

Slide 17:

Spark集群部署模式中，哪种方式是Spark自带的集群管理器？

正确答案 / 정답: C

Slide 18:

以下哪个不是Spark的核心组件？

正确答案 / 정답: D

Slide 19:

Spark与Mesos之间性能匹配好的主要原因是什么？

/ 모두 메모리 컴퓨팅 기반임

正确答案 / 정답: B

Slide 20:

Spark 2.3版本引入的新功能是什么？

正确答案 / 정답: C

Slide 21:

在YARN管理的Spark生态系统中，底层使用的分布式文件系统是什么？

正确答案 / 정답: A

Slide 22: 마무리

좋습니다. 오늘 Spark의 배포와 응용 방식에 대한 학습은 여기까지입니다.

여러분이 Spark의 다양한 환경에서의 배포 방법과 각 배포 방식의 장단점을 마스터하시기를 바랍니다.

경청해주셔서 감사합니다!

Slide 22: 감사 인사

好的，今天关于Spark的部署和应用方式的学习就到这里。

希望大家能够掌握Spark在不同环境下的部署方法，以及各种部署方式的优缺点。

感谢大家的聆听！