УНИВЕРСИТЕТ ИТМО

Технологии хранения больших данных

Лекция №6. Форматы хранения данных – таблицы, protobuf, avro

Ирина Алексеевна Радченко
iradche@gmail.com

Введение

Обзор темы

- Значение форматов хранения данных
- Различные форматы: Таблицы, Protocol Buffers (Protobuf), Avro

Цели лекции

- Понимание основных форматов хранения данных
- Изучение преимуществ и недостатков каждого формата

Таблицы

Определение

- Структурированный формат данных, представленный в виде строк и столбцов

Примеры использования

- Реляционные базы данных, CSV-файлы

Преимущества

- Простота понимания и использования
- Легкость интеграции с различными системами

Пример таблицы

ID	Name	Age
1	John Doe	29
2	Jane Smith	34
3	Emily Davis	22

Protocol Buffers (Protobuf)

Определение

- Формат сериализации данных, разработанный Google

Особенности

- Высокая производительность и компактность
- Независимость от языка программирования

Примеры использования

- Взаимодействие микросервисов, хранение конфигурационных данных

Пример структуры Protobuf

```
syntax = "proto3";

message Person {
  int32 id = 1;
  string name = 2;
  int32 age = 3;
}
```

Avro

Определение

- Формат данных для сериализации, разработанный Apache

Особенности

- Поддержка схем для валидации данных
- Оптимизирован для Hadoop

Примеры использования

- Хранение и передача данных в экосистеме Hadoop

Пример схемы Avro

```
{
 "type": "record",
 "name": "Person",
 "fields": [
  {"name": "id", "type": "int"},
  {"name": "name", "type": "string"},
  {"name": "age", "type": "int"}
 ]
}
```

Сравнение форматов хранения данных

Таблицы vs. Protobuf vs. Avro

- Простота использования, производительность, объем занимаемого пространства

Критерии выбора

- Тип данных, требования к производительности, совместимость с системами

Преимущества и недостатки таблиц

Преимущества

- Легкость чтения и редактирования
- Широкая поддержка инструментов анализа

Недостатки

- Низкая эффективность при больших объемах данных
- Ограниченная производительность для сложных структур

Преимущества и недостатки Protobuf

Преимущества

- Высокая производительность, компактность
- Независимость от языка

Недостатки

- Необходимость компиляции схемы
- Ограниченная поддержка сложных типов данных

Преимущества и недостатки Avro

Преимущества

- Поддержка динамических схем
- Оптимизация для экосистемы Hadoop

Недостатки

- Требуется знание схемы для работы с данными
- Сложность настройки

Примеры использования в реальных системах

Таблицы

- Анализ данных, бизнес-отчеты

Protobuf

- Взаимодействие микросервисов, хранение конфигураций

Avro

- Анализ данных в Hadoop, потоковая передача данных

Практические примеры

Реальные примеры использования различных форматов данных

- Сценарии использования, сравнение производительности

Заключение

Подведение итогов

- Важность выбора подходящего формата для конкретной задачи
- Влияние формата данных на производительность и эффективность системы

Дополнительные материалы и литература

Вопросы и обсуждение

Вопросы от студентов

Обсуждение примеров

Благодарность за внимание

Контактная информация

- Email: iradche@gmail.com
- Телеграм: @dadaistka

Следующая лекция

- Тема следующей лекции:

  «Озера данных, витрины данных»

Спасибо за внимание!

www.ifmo.ru

ITMO more than a UNIVERSITY