VILNIAUS GEDIMINO TECHNIKOS UNIVERSITETAS

Didieji duomenys dirbtinio intelekto kontekste

Andrej Bugajev

2020 m. gruodžio 15 d.

Didieji duomenys

Apibėžimas

- Jeigu duomenų daug – duomenys didieji?

Didieji duomenys

Apibėžimas

- Jeigu duomenų daug – duomenys didieji?
- Duomenų patekimą į didžiųjų duomenų kategoriją, paprastai sąlygoja *Apimtis*, *Įvairovė*, *Greitis* (angl. 3 V's: Volume, Variety, Velocity).

Didieji duomenys

Apibėžimas

- Jeigu duomenų daug – duomenys didieji?
- Duomenų patekimą į didžiųjų duomenų kategoriją, paprastai sąlygoja *Apimtis*, *Įvairovė*, *Greitis* (angl. 3 V's: Volume, Variety, Velocity).
- Didžiųjų duomenų termino atsiradimą sąlygojo tam tikrų (didžiųjų duomenų) technologijų taikymas uždaviniams, kuriems tos technologijos gerai tinka.

Google File System (GFS)

GFS sukurta atsižvelgiant į šias savybes

- Gedimai įvyksta dažnai (žemos klasės skaičiuojamoji įranga) – žemo patikimumo "geležies" naudojimo palaikymas
- Failai yra dideli (GB–TB eilės)
- Vyrauja didelės apimties nuoseklieji įrašymai, kurie pildo failus jų neperrašant iš naujo (angl. append).
- Failų operacijų laikas pagrinde yra ribojamas duomenų pralaidumu bet ne uždelsimais tarp operacijų.

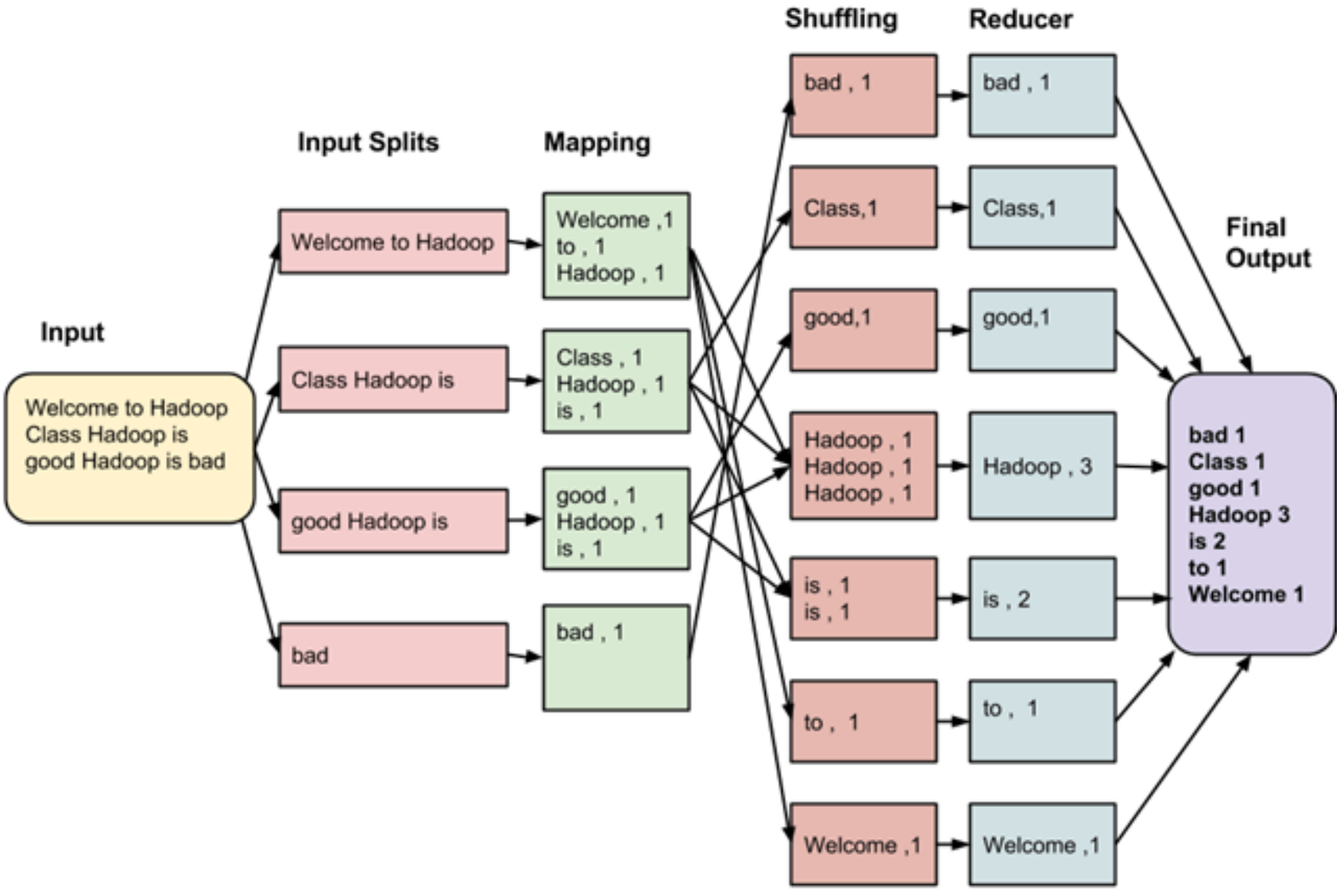
GFS netinka, kai:

- Yra mažo uždelsimo duomenų nuskaitymo poreikis. Optimizuotas pralaidumas nepaisans uždelsimo.
- Daug smulkių failų.
- Yra nuolatinis duomenų kitimas. Sistemoje optimizuotas didelių duomenų masyvų nuoseklus įrašymas.

Lentelė: Labiausiai paplitusios didžiųjų duomenų sistemos, sugrupuotos pagal programavimo modelius.

	Sistema	Abstrakcijos lygmuo	Lygiagretumo tipas	Infrastruktūros mastas	Taikymo klasės
MapReduce	Apache Hadoop	Žemas	Data	Didelis	General purpose
DAG	Apache Spark	Žemas	Data/Task	Didelis	General purpose
	Apache Storm	Vidutinis	Data/Task/Pipeline	Didelis	Real-time stream processing
	Apache Flink	Vidutinis	Data/Task/Pipeline	Didelis	Real-time stream processing
	Azure ML	Aukštas	Task	Mažas	Predictive analytics and machine learning
MP	MPI	Žemas	Data	Vidutinis	General purpose
BSP	Apache Giraph	Žemas	Data	Mažas	Graph processing
	Apache Hama	Žemas	Data	Mažas	Graph processing
Workflow	Swift	Vidutinis	Task/Pipeline	Vidutinis	Scientific data analytics
	COMPSs	Vidutinis	Task	Mažas	Scientific data analytics
	DMCF	Aukštas	Task/Pipeline	Mažas	Visual and script-based analytics
SQL-like	Apache Pig	Vidutinis	Data/Task	Didelis	Data querying and analysis
	Apache Hive	Aukštas	Data	Didelis	Data querying and reporting

MapReduce


Input
Welcome to Hadoop
Class Hadoop is
good Hadoop is bad
Input Splits
Welcome to Hadoop
Class Hadoop is
good Hadoop is
bad
Mapping
Welcome ,1
to , 1
Hadoop , 1
Class , 1
Hadoop , 1
is , 1
good , 1
Hadoop , 1
is , 1
bad , 1
Shuffling
bad , 1
Class,1
good,1
Hadoop , 1
Hadoop , 1
Hadoop , 1
is , 1
is , 1
to , 1
Welcome ,1
Reducer
bad , 1
Class,1
good,1
Hadoop , 3
is , 2
to , 1
Welcome ,1
Final
Output
bad 1
Class 1
good 1
Hadoop 3
is 2
to 1
Welcome 1

Apache Spark

6 Apache Spark ekosistemos komponentai

- Apache Spark-Spark Core,
- Spark SQL,
- Spark Streaming,
- Spark MLlib,
- Spark GraphX,
- SparkR