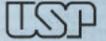
## Análise de Dados com Base em Processamento Massivo em Paralelo

# Aula 8: Explorando o Módulo pyspark.sql

Cristina Dutra de Aguiar ICMC/USP cdac@icmc.usp.br







#### Conteúdo das Aulas

	Foco	Detalhamento
Aula 05	Pandas	diversos métodos consultas sobre DataFrames
Aula 06	Apache Spark RDD	diversos métodos manipulação de RDDs
Aula 07	Módulo pyspark.sql	método spark.sql() consultas textuais em SQL
Aula 08	Módulo pyspark.sql	diversos métodos consultas sobre DataFrames







# Infraestrutura Computacional

	Processamento	Ambiente
Aula 05 Pandas	centralizado	um computador
Aula 06 Spark RDD		<i>cluster</i> de computadores
Aula 07 spark.sql()	paralelo e distribuído	ou
Aula 08 pyspark.sql		computação em nuvem







#### **Processamento dos Dados**

	Persistência em Disco	Computação em RAM
Aula 05 Pandas	sistema de arquivos local	centralizada
Aula 06 Spark RDD		
Aula 07 spark.sql()	sistema de arquivos distribuído (HDFS)	distribuída
Aula 08 pyspark.sql		





## Complexidade de Instalação

	Módulos	Outras Configurações
Aula 05 Pandas	Pandas com suporte de Python	-
Aula 06 Spark RDD	Java	criação e configuração
Aula 07 spark.sql()	Spark findspark	de sessão configuração de variáveis de ambiente
Aula 08 pyspark.sql	pyspark	GC GITIDICITE







## Volume de Dados

	Volume	Escalabilidade Atual
Aula 05 Pandas	grandes volumes	até 5GB
Aula 06 Spark RDD		
Aula 07 spark.sql()	gigantescos volumes	até 1TB
Aula 08 pyspark.sql		





### Velocidade de Dados

	Latência	Módulo
Aula 05 Pandas	lote	-
Aula 06 Spark RDD	lote e streaming	Apache Spark Streaming
Aula 07 spark.sql()	lote	Apache Spark SQL
Aula 08 pyspark.sql		Apacile Spain SQL



### Variedade de Dados

	Tipo de Dados	Conceito Subjacente
Aula 05 Pandas	estruturados	DataFrames
Aula 06 Spark RDD	estruturados, semiestruturados e não estruturados	RDDs
Aula 07 spark.sql()	estruturados	DataFrames
Aula 08 pyspark.sql		construídos sobre RDDs







## Abstração do Conceito de RDD

	Nível	Detalhamento
Aula 05 Pandas	-	-
Aula 06 Spark RDD	baixo	RDDs manipulados diretamente
Aula 07 spark.sql()	alto	uso de comandos SQL
Aula 08 pyspark.sql	alto	uso de métodos funcionais







# Abstração Problema e Solução

	Programação	Detalhamento
Aula 05 Pandas	funcional	como os dados
Aula 06 Spark RDD		devem ser obtidos
Aula 07 spark.sql()	declarativa	quais dados devem ser obtidos
Aula 08 pyspark.sql	funcional	como os dados devem ser obtidos

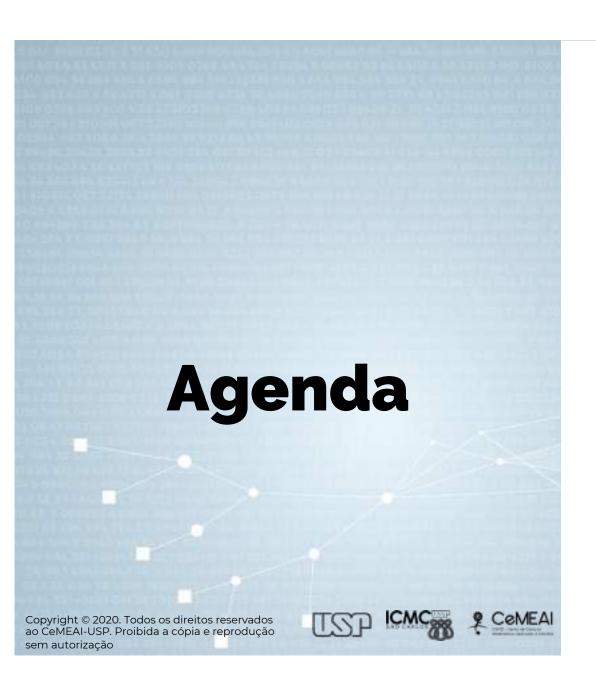


#### Grau de Conhecimento do Usuário

	Experiente em	
Aula 05 Pandas	resolução de consultas passo a passo	
Aula 06 Spark RDD	uso de métodos de baixo nível programação paralela e distribuída	
Aula 07 spark.sql()	uso da linguagem SQL	
Aula 08 pyspark.sql	resolução de consultas passo a passo programação paralela e distribuída	







- Métodos de Interesse
- Consultas OLAP
- Comparativo Pandas, Spark
  RDD e Spark SQL