Spark Summit North America 202006 高清 PPT 下载

为期五天的 Spark Summit North America 2020在美国时间 2020-06-22 ~ 06-26 举行。由于今年新冠肺炎的影响，本次会议第一次以线上的形式进行。这次会议虽然是五天，但是前两天是培训，后面三天才是正式会议。本次会议一共有超过210个议题，一如既往，主题也主要是 Spark + AI，在 AI 方面会议还深入讨论一些流行的软件框架，如 Delta Lake、MLflow、TensorFlow、SciKit-Learn、Keras、PyTorch、DeepLearning4J、BigDL 和 deep learning pipeline等。会议的全部日程请参见：https://databricks.com/sparkaisummit/north-america-2020/agenda

这次会议带来了几点比较重要消息：数砖收购 Redash 公司，发布 Delta Engine等，不过目前 KeyNote 会议的 PPT 还没有发布，感兴趣的可以看下相关视频。过往记忆大数据也在前几天发了几篇这次会议 KeyNote 的介绍，感兴趣的同学可以看这里。另外，在接下来的几天，本公众号也会对一些比较有意思的议题进行介绍，敬请关注本公众号。

如果想及时了解Spark、Hadoop或者HBase相关的文章，欢迎关注微信公众号：iteblog_hadoop

本次会议的议题范围具体如下：

- Apache Spark™, Delta Lake, MLflow 以及 Koalas 未来规划；
- 管理机器学习生命周期的最佳实践
- 构建大规模可靠数据管道的技巧
- 流行的深度学习和机器学习框架的最新发展
- 真实的 AI 用户案例

下载途径

关注微信公众号 过往记忆大数据 或者 Java技术范 并回复 spark-9832 获取。

可下载的PPT

下面议题提供 PPT 下载

- Data Science Across Data Sources with Apache Arrow
- Portable Scalable Data Visualization Techniques for Apache Spark and Python Notebook-based Analytics
- Native Support of Prometheus Monitoring in Apache Spark 3.0
- Performant Streaming in Production: Preventing Common Pitfalls when Productionizing Streaming Jobs
- Scaling Security Threat Detection with Apache Spark and Databricks
- User Defined Aggregation in Apache Spark: A Love Story
- Powering Interactive BI Analytics with Presto and Delta Lake
- Using AI to Support Proliferating Merchant Changes
- Tuning ML Models: Scaling, Workflows, and Architecture
- Battling Model Decay with Deep Learning and Gamification
- An Approach to Data Quality for Netflix Personalization Systems
- High-Performance Analytics with Probabilistic Data Structures: the Power of HyperLogLog
- Preventing Abuse Using Unsupervised Learning
- Geospatial Analytics at Scale: Analyzing Human Movement Patterns During a Pandemic
- Leveraging Apache Spark for Scalable Data Prep and Inference in Deep Learning
- Filtering vs Enriching Data in Apache Spark
- Scalable Acceleration of XGBoost Training on Apache Spark GPU Clusters
- Deep Dive into GPU Support in Apache Spark 3.x
- Sputnik: Airbnb's Apache Spark Framework for Data Engineering
- Patterns and Anti-Patterns for Memorializing Data Science Project Artifacts
- Automated and Explainable Deep Learning for Clinical Language Understanding at Roche
- Building Understanding Out of Incomplete and Biased Datasets using Machine Learning and Databricks
- Encryption and Masking for Sensitive Apache Spark Analytics Addressing CCPA and Governance
- Managing ADLS gen2 using Apache Spark
- Using Apache Spark and Differential Privacy for Protecting the Privacy of the 2020 Census Respondents
- The 2020 Census and Innovation in Surveys
- scaling-data-and-ml-with-apache-spark-and-feast
- The Apache Spark File Format Ecosystem
- Building the Petcare Data Platform using Delta Lake and 'Kyte': Our Spark ETL Pipeline
- A Production Quality Sketching Library for the Analysis of Big Data
- Children Safety Retrieval (CENSER) System for Retrieval of Kidnapped Children from Brothels in India

- Benchmark Tests and How-Tos of Convolutional Neural Network on HorovodRunner Enabled Apache Spark Clusters
- Scalable AutoML for Time Series Forecasting using Ray
- Using Machine Learning to Evolve Sports Entertainment
- Using Bayesian Generative Models with Apache Spark to Solve Entity Resolution Problems (DeDup, Merging, Uniqueness) at Scale
- Fine Tuning and Enhancing Performance of Apache Spark Jobs
- All In - Migrating a Genomics Pipeline from BASH/Hive to Spark (Azure Databricks) - A Real World Case Study
- Running Apache Spark on Kubernetes: Best Practices and Pitfalls
- Lessons Learned from Modernizing USCIS Data Analytics Platform
- On Improving Broadcast Joins in Apache Spark SQL
- Using Databricks as an Analysis Platform
- Is This Thing On? A Well State Model for the People
- Advanced Natural Language Processing with Apache Spark NLP
- Building a Streaming Microservice Architecture: with Apache Spark Structured Streaming and Friends
- Simplify and Boost Spark 3 Deployments with Hypervisor-Native Kubernetes
- Deploying Apache Spark Jobs on Kubernetes with Helm and Spark Operator
- Resource-Efficient Deep Learning Model Selection on Apache Spark
- Bring Satellite and Drone Imagery into your Data Science Workflows
- Scoring at Scale: Generating Follow Recommendations for Over 690 Million LinkedIn Members
- From HDFS to S3: Migrate Pinterest Apache Spark Clusters
- SparkCruise: Automatic Computation Reuse in Apache Spark
- Chromatic Sparse Learning
- Deploy and Serve Model from Azure Databricks onto Azure Machine Learning
- Cloud-Native Apache Spark Scheduling with YuniKorn Scheduler
- The Revolution Will be Streamed
- Democratizing PySpark for Mobile Game Publishing
- Ray: Enterprise-Grade, Distributed Python
- Fugue: Unifying Spark and Non-Spark Ecosystems for Big Data Analytics
- Enabling Scalable Data Science Pipeline with Mlflow at Thermo Fisher Scientific
- Scaling Up AI Research to Production with PyTorch and MLFlow
- Best Practices for Building Robust Data Platform with Apache Spark and Delta
- Building a Pipeline for State-of-the-Art Natural Language Processing Using Hugging Face Tools
- Designing the Next Generation of Data Pipelines at Zillow with Apache Spark
- Lessons from Building Large-Scale, Multi-Cloud, SaaS Software at Databricks
- Flash for Apache Spark Shuffle with Cosco
- Building a Real-Time Feature Store at iFood
- AutoML Toolkit – Deep Dive
- Operationalize Apache Spark Analytics
- End-to-End Deep Learning with Horovod on Apache Spark
- Building Data Quality Audit Framework using Delta Lake at Cerner
- Zipline - A Declarative Feature Engineering Framework

- Automating Federal Aviation Administration's (FAA) System Wide Information Management (SWIM) Data Ingestion and Analysis
- Apache Spark NLP for Healthcare: Lessons Learned Building Real-World Healthcare AI Systems
- A Thorough Comparison of Delta Lake, Iceberg and Hudi
- Productionizing Machine Learning Pipelines with Databricks and Azure ML
- Advertising Fraud Detection at Scale at T-Mobile
- AI-Assisted Feature Selection for Big Data Modeling
- The Data Lake Engine Data Microservices in Spark using Apache Arrow Flight
- Ibis: Seamless Transition Between Pandas and Apache Spark
- Simplify CDC Pipeline with Spark Streaming SQL and Delta Lake
- Power of Visualizing Embeddings
- Deliver Dynamic Customer Journey Orchestration at Scale
- Top Down Specialization Using Apache Spark
- The Killer Feature Store: Orchestrating Spark ML Pipelines and MLflow for Production
- Tackling Scaling Challenges of Apache Spark at LinkedIn
- Scaling up Deep Learning by Scaling Down
- Wood Log Inventory Estimation using Image Processing and Deep Learning Technique
- Building Identity Graphs over Heterogeneous Data
- Productionizing Machine Learning with Apache Spark, MLflow and ONNX from the ground to cloud using SQL Server
- Generative Hyperloop Design: Managing Massively Scaled Simulations Focused on Quick-Insight Analytics and Demand Modelling
- Efficiently Building Machine Learning Models for Predictive Maintenance in the Oil & Gas Industry with Databricks
- Consolidate Your Technical Debt With Spark Data Sources -Tools and Techniques to Integrate Native Code
- Running Emerging AI Applications on Big Data Platforms with Ray On Apache Spark
- Best Practices for Engineering Production-Ready Software with Apache Spark
- Automatic Forecasting using Prophet, Databricks, Delta Lake and MLflow
- Composable Data Processing with Apache Spark
- Accelerating Spark SQL Workloads to 50X Performance with Apache Arrow-Based FPGA Accelerators
- Accelerating the ML Lifecycle with an Enterprise-Grade Feature Store
- Faster Data Integration Pipeline Execution using Spark-Jobserver
- Continuous Delivery of ML-Enabled Pipelines on Databricks using MLflow
- Accelerating Apache Spark Shuffle for Data Analytics on the Cloud with Remote Persistent Memory Pools
- Bucketing 2.0: Improve Spark SQL Performance by Removing Shuffle
- How to Performance-Tune Apache Spark Applications in Large Clusters
- Saving Energy in Homes with a Unified Approach to Data and AI
- Productionizing Deep Reinforcement Learning with Spark and MLflow
- SQL Performance Improvements at a Glance in Apache Spark 3.0
- Pandas UDF and Python Type Hint in Apache Spark 3.0
- Parallelization of Structured Streaming Jobs Using Delta Lake
- Artificial Lawyers. Will Your Next Attorney be a Machine?

- Adaptive Query Execution: Speeding Up Spark SQL at Runtime
- How Azure and Databricks Enabled a Personalized Experience for Customers and Patients at CVS Health
- Optimize the Large Scale Graph Applications by using Apache Spark with 4-5x Performance Improvements
- Data Mesh in Practice: How Europe's Leading Online Platform for Fashion Goes Beyond the Data Lake
- Running Apache Spark Jobs Using Kubernetes
- Koalas: Making an Easy Transition from Pandas to Apache Spark
- Vectorized Deep Learning Acceleration from Preprocessing to Inference and Training on Apache Spark in SK Telecom
- Text Extraction from Product Images Using State-of-the-Art Deep Learning Techniques
- Care and Feeding of Catalyst Optimizer
- Apache Spark vs Apache Spark: An On-Prem Comparison of Databricks and Open-Source Spark
- Enabling Physics and Empirical-Based Algorithms with Spark Using the Integration of MATLAB in Databricks
- Democratizing Data
- Evolution is Continuous, and so are Big Data and Streaming Pipelines
- Geospatial Options in Apache Spark
- Scaling Production Machine Learning Pipelines with Databricks
- Taming the Search: A Practical Way of Enforcing GDPR and CCPA in Very Large Datasets with Apache Spark
- Zeus: Uber's Highly Scalable and Distributed Shuffle as a Service
- Productionizing Machine Learning with a Microservices Architecture
- Productionalizing Models through CI/CD Design with MLflow
- DataSource V2 and Cassandra – A Whole New World
- Hyperspace: An Indexing Subsystem for Apache Spark
- Data Driven Decisions at Scale
- Deep Dive into the New Features of Apache Spark 3.0
- Securing Apache Spark Applications at Facebook
- Building a Feature Store around Dataframes and Apache Spark
- Tracing the Breadcrumbs: Apache Spark Workload Diagnostics
- Enabling Push Button Productization of AI Models
- Everyday Probabilistic Data Structures for Humans
- Deep Learning Enabled Price Action with Databricks and AWS
- Clinical Suspecting at Scale Using PySpark
- Using Apache Spark for Predicting Degrading and Failing Parts in Aviation
- Operationalizing Big Data Pipelines At Scale
- Columbia Migrates from Legacy Data Warehouse to an Open Data Platform with Delta Lake
- How Adobe Does 2 Million Records Per Second Using Apache Spark!
- Accelerating Data Processing in Spark SQL with Pandas UDFs
- Building a Federated Data Directory Platform for Public Health
- Translating Models to Medicine an Example of Managing Visual Communications
- Delta from a Data Engineer's Perspective

- Disrupting Risk Management through Emerging Technologies
- Automated Testing For Protecting Data Pipelines from Undocumented Assumptions
- Geosp.AI.tial: Applying Big Data and Machine Learning to Solve the World's Toughest Geospatial Intelligence Problems
- Delta from a Data Engineer's Perspective
- Healthcare Claim Reimbursement using Apache Spark
- From Idea to Model: Productionizing Data Pipelines with Apache Airflow
- Willump: Optimizing Feature Computation in ML Inference
- Real-Time Forecasting at Scale using Delta Lake and Delta Caching
- Leveraging Apache Spark to Develop AI-Enabled Products and Services at Bosch
- Continuous Delivery of Deep Transformer-Based NLP Models Using MLflow and AWS Sagemaker for Enterprise AI Scenarios
- From Python to PySpark and Back Again – Unifying Single-host and Distributed Deep Learning with Maggy
- Shparkley: Scaling Shapley with Apache Spark
- Understanding and Improving Code Generation
- User Defined Aggregation in Apache Spark: A Love Story
- Machine Learning Data Lineage with MLflow and Delta Lake
- Memory Optimization and Reliable Metrics in ML Pipelines at Netflix
- Operationalizing Machine Learning at Scale at Starbucks
- Presto on Apache Spark: A Tale of Two Computation Engines
- Generalized SEIR Model on Large Networks
- Deep Learning at Scale with Apache Spark and Determined
- How R Developers Can Build and Share Data and AI Applications that Scale with Databricks and RStudio Connect
- Rapid Response to Hospital Operations using Data and AI during COVID-19