DATA PLATFORMS LIVE 2018

April 11-13, 2018 - The Wigwam Resort, Phoenix, AZ

SPONSORED BY Qubole

Session Guide

Wednesday, April 11		Location
9:00 am - 6:00 pm	**Registration & Networking**	
9:30am - 11:30 am	**Training: Qubole Enterprise User** This session will answer the following questions: What is Qubole? What are the features available to me as a user? How does it interact with the cloud on my behalf? How do I pick the appropriate SQL Engine for my need?	Mohave West
11:30 am - 1:00 pm	**Lunch Break** (on own)	
1:00 pm - 3:00 pm	**Training: Qubole Enterprise Admin (AWS)** This session will address how Qubole clusters work, how to administer Qubole clusters, and how to decide which cluster is appropriate for a given scenario. This session will be specific to AWS.	Mohave West
1:00 pm - 3:00 pm	**Training: Qubole Enterprise (Azure)** This session will address how Qubole clusters work, how to administer Qubole clusters, and how to decide which cluster is appropriate for a given scenario. This session will be specific to Azure.	Mohave East
3:30 pm - 5:30 pm	**Training: Spark for Data Scientists** This presentation will start with a summary of data science and its importance by Piero Cinquegrana, Qubole's Data Science Product Manager. Subsequently, Alex Aidun will review the features in Qubole that support data scientists during their development cycles - topics include Spark Notebooks, Qubole Features, Notebook API Execution, Notebook Dashboards, and Notebook Interpreter Configuration.	Mohave West
6:30 pm - 8:30 pm	**Welcome Reception** Get your fill of fun, food, and drinks as we launch Data Platforms 2018. Join your fellow attendees for a cantina-style kickoff in one of The Wigwam's lounges evoking old-world charm.	Wigwam Foyer & Ballroom

Thursday, April 12		Location
7:30 am - 8:45 am	**Breakfast**	Sachem
9:00am - 9:45 am	**Opening Keynote: Big Data Activation** *Ashish Thusoo, Co-Founder & CEO, Qubole* This keynote will discuss the gap that enterprises face today when activating their big data. It will make a case for the shift that organizations need to make towards a big data activation strategy in order to put their data assets to use for differentiating and achieving business objectives. The session will also cover key elements of big data activation supported by usage trends of Qubole's cloud-native big data activation platform. It will present various ways that enterprises can use to measure their own activation readiness, and demonstrate why Qubole provides the right approach to big data activation.	Wigwam Ballroom
9:45am - 10:15 am	**Panel: Activating Big Data Across the Enterprise** Every CEO aspires to create a data-driven culture that can activate 100s or 1000s of users and petabyte-scale data to continuously deliver true business value. This keynote panel will explore the journey of 4 companies: Comcast, Turner Broadcasting, Fanatics and MediaMath, that have chronicled their successes and challenges in two books by O'Reilly Media about Creating a Data-Driven Enterprises. The panelists will talk not just about their technology strategy and choices but also how data-driven insights are powering their business and transforming the competitive dynamics of their industry. **Panelists** • Barbara Eckman, Principal Data Architect, Comcast • Santanu Dey, Director of Data Science & Engineering, Fanatics • John Slocum, Vice President, MediaMath Data Management Platform • Vikram Marathe, Technical Director of Development and Architecture, Turner Data Cloud **Moderated by:** Jose Villacis, Senior Director, Product Marketing	Wigwam Ballroom
10:30 am - 12:15 pm	**Tech Talks** Practitioners share best practices, techniques, challenges, and solutions in these rapid-fire sessions on the technical, organizational, and cultural aspects of building a modern big data platform. We'll run five sessions per hour from 10:30 am - 12:15 pm; choose one session per hour.	

Session I	10:30 am – 11:15 am	Location

HiveServer2 Or: How I Learned to Stop Worrying and Love the Bomb — Mohave East

Puneet Jaiswal, Software Engineer, Data Platforms and Infrastructure, Lyft

This presentation covers Apache Hive use cases at Lyft, including the challenges and learnings associated with the recent Hive upgrade and handling ETL at scale.

Building a Real-Time Decision Engine Using Machine Learning on Apache Spark Structured Streaming — Aztec A

Garren Staubli, Senior Data Engineer, Blueprint Technologies

Real-time decision making using machine learning (ML) / artificial intelligence (AI) is the holy grail of customer-facing applications. It's no longer a long-shot dream; it's our new reality.

The real-time decision engine leverages the latest features in Apache Spark 2.3, including stream-to-stream joins and Spark ML, to directly improve the customer experience. We will discuss the architecture at length, including data source features and technical intricacies, as well as model training and serving dynamics. Critically, real-time decision engines that directly affect customer experience require production-level SLAs and/or reliable fallbacks to avoid meltdowns, which this talk will also address.

A Lap Around Azure Data Lake — Mohave West

Francesco Diaz, Regional Solutions Manager Alps, Nordic & Southern Europe, Insight

Azure Data Lake is one of the most powerful PaaS services that Microsoft Azure offers to manage Big Data. Built on well-known projects such as HDFS and YARN, it allows for the ability to focus on the design of the solution instead of the administration part. A new language, U-SQL, combines SQL and C# to work with any type and size of data. During the session, we will explore Azure Data Lake Store and Azure Data Lake Analytics, the core components of the Azure Data Lake offering.

The BIG Picture: The Journey from On-Premise to Cloud to the ML Marketplace for a Geospatial Data Platform — Aztec B

Ori Elkin, Chief Product Officer, DigitalGlobe

In a world driven by location intelligence, DigitalGlobe is creating a place where everyone can access geospatial data and use it to derive truth and, in turn, knowledge. DigitalGlobe's Geospatial Big Data Platform (GBDX) is empowering an ecosystem of location intelligence, cataloging hundreds of petabytes worth of geospatial information and executing tens of millions (!) of hours' worth of cloud compute.

This session will talk about the migration of GBDX from on-premise to cloud, involving 100 petabytes of satellite image data. We will also discuss GBDX as a key platform for analytics and machine learning applications across industries, and how it's evolving into a marketplace for imagery-related machine learning algorithms.

The Story of Building a Scalable Data Trust Playbook at Optimizely — Aztec C

Vignesh Sukumar, Senior Manager, Data Engineering, Optimizely

At Optimizely, we receive billions of user click stream events for the thousands of A/B experiments we run for our customers every day. Previously, customer inquiries related to the alignment of key experiment metrics between raw data and experiment results required expensive engineering analysis due to lack of scale and flexibility. In this talk, I will walk the audience through the journey of how we created a playbook to enhance customer trust in these situations and make self-service scalable for the entire organization.

Session II	11:30 am – 12:15 pm	Location

Packaging, Deploying, and Running Apache Spark Applications in Production

Mohave East

Saba El-Hilo, Data Engineer, Mapbox

Apache Spark has proven to be indispensable due to its endless applications and use cases. Developers, data scientists, engineers, and analysts alike can benefit from its power. However, deterministically managing dependencies, packaging, testing, scheduling, and deploying a Spark application can be challenging.

As organizations grow, these individuals become dispersed across multiple teams and departments. This makes a team-specific solution no longer applicable. So, what type of tooling do you need to allow these individuals to solely focus on writing a Spark application? And more importantly, how do you enforce development best practices such as unit testing, continuous integration, version control, and deployment environments?

The data engineering team at Mapbox has developed tooling and infrastructure to address these challenges and enable individuals across the organization to build and deploy Spark applications. This talk will walk you through our solution to packaging, deploying, scheduling, and running Spark applications in production. We will also address some of the problems we've faced and how the adoption process is evolving across the team.

From Zero to Activated Big Data in the Cloud – The First Year's Journey

Aztec A

Brian Greene, Cloud Data Architect, Auris Surgical Robotics

What would you do in this scenario: you have a blank slate, one year to prepare for "big data is on the way," and your company's acknowledgement that data is a strategic corporate asset?

Attend this session to explore the goals, best practices, architectural constraints, and technologies that shaped the journey from that starting point to continuously delivered and live systems with engaged users. Also hear about the multiple use cases downstream from the data lake, such as APIs, guided exploration, streaming, and integration with external systems. Finally, learn how to accomplish all of this with one full-time employee and strategic partnerships.

Presto: Fast SQL on Everything

Mohave West

David Phillips, Software Engineer, Facebook

Presto is an open source distributed query engine that supports much of the SQL analytics workload at Facebook. This talk introduces a selection of Facebook use cases, which range from user-facing reporting applications to multi-hour ETL jobs, then explains the architecture, implementation, features, and performance optimizations that enable Presto to support these use cases.

	Lighthouse Related Product, an Efficient Cross-Boundary Product Recommendation Platform on Qubole Ecosystem *Jing Pan, User Experience Researcher, Fanatics* Fanatics, Inc. will introduce an item-to-item recommendation service platform in production, Lighthouse Related Product (LRP). LRP offers supervised-versus-non-supervised boundary-crossing and is extendable, flexible, and lightweight. LRP implements a modeling architecture that fuses into one system heterogeneous features from modern machine learning techniques of (1) non-supervised user-item matrices, (2) self-supervised Word2Vec, and (3) supervised XGBoost or deep learning. This architecture allows innate extendibility to user-item recommendations, and flexibility for both offline and online use cases. It is lightweight and efficient enough to handle near one million products' item-to-item recommendations on over 400 affiliated sites. The platform relies on the Apache Spark cluster in Qubole for both data feature extraction and prediction in a distributed manner with map procedure from a pre-trained supervised model; tasks on the Spark cluster in Qubole are seamlessly integrated into the rest of the workflows in Fanatics with another third-party scheduling service, Stone Branch. LRP has successfully passed the real-life load test of the 2017 holiday season and Super Bowl LII, and an earlier predecessor of the current version of LRP had achieved better performance in all measures, such as click-through rate and average order volume, compared to an industrial standard third-party recommendation service provider.	Aztec B
	A Framework for Assessing the Quality of Product Usage Data *David Oh, Data Engineer for the Autodesk Data Platform, Autodesk* This presentation will discuss the importance of data quality and outline an approach to assess and measure the quality of product usage event logs. A data quality assessment framework helps build trust in our data and enables analysts to generate a deep understanding of product usage patterns, product stability, and utilization of purchased assets by our customers. Unlocking valuable insights from this data depends on the presence of high-quality and complete data sets that provide the ability to link product usage events with back office accounts and entitlement data.	Aztec C
12:15 – 1:15 pm	**Lunch**	Sachem
1:30 - 2:15 pm	**IBM Keynote: Demystifying AI, Machine Learning & Deep Learning** *Sumit Gupta, Vice President, Artificial Intelligence, Machine Learning and HPC, IBM Cognitive Services* From chat bots and recommendation engines to Google Voice and Apple Siri, artificial intelligence (AI) has begun to permeate our lives. We will demystify AI, present the difference between machine learning and deep learning, share why the huge interest is occurring now, show some fun use cases and demos, and discuss use cases of how deep learning–based AI methods can be used to garner insights from enterprise data. We will also talk about what IBM is doing to make deep learning and machine learning more accessible and useful to a broader set of data scientists.	Wigwam Ballroom
2:30 – 5:15 pm	**Tech Talks continue** We'll run five sessions per hour from 2:30 pm - 5:15 pm; choose one session per hour.	

Session III	2:30 pm – 3:15 pm	Location

Self-Regulating Streaming Capabilities in Apache Heron

Mohave East

Karthik Ramasamy, CEO & Co-founder, Streamlio

Several enterprises have been producing data not only at high volume but also at high velocity. Many daily business operations depend on real-time insights, and therefore real-time staging and processing of the data is gaining significance. Thus, there is a need for a scalable infrastructure that can continuously ingest and process billions of events per day the moment the data is acquired.

To achieve real-time performance at scale, Twitter designed Heron for stream data processing. In production for more than four years, Heron faced crucial challenges from an operational point of view: the manual, time-consuming, and error-prone tasks of tuning various configuration knobs to achieve service level objectives (SLO), as well as the maintenance of SLOs in the face of sudden, unpredictable load variation and hardware or software performance degradation.

In order to address these issues, we conceived and implemented several innovative methods and algorithms that aim to bring self-regulating capabilities to these systems, thereby reducing the number of manual interventions. In this talk, we will give a brief introduction to issues and enumerate challenges such as slow hosts, unpredictable spikes, network slowness, and network partitioning that we faced in production. We'll also describe how we made the systems self-regulating to minimize overhead and operations.

Key Objectives and Principles for Building Predictive Models on Big Data

Aztec A

Satya Ramachandran, Vice President, Engineering, Neustar

Business analysts spend a lot of time today looking at what happened in the past, but what about trying to grasp what will happen in the future? For example, what if you are given 10 percent more budget for next quarter's marketing spend? Do you know how you'll use that extra money, and do you know what impact it will create? Or suppose you want to increase your budget, but need to show what you expect that increase to do - then what?

Many of today's data applications are simply "decision support systems" designed to be useful in the aforementioned scenarios. They help business professionals use data to better understand their environment and make better decisions. But with larger volumes of data and increased ambitions of competitive businesses, the end goals become tougher to achieve. As the VP of Engineering for MarketShare DecisionCloud at Neustar, which provides planning and analytics capabilities for marketers, Satya Ramachandran has taken on these challenges by leveraging big data technologies.

In this talk, Satya will discuss some of the high expectations he's faced at MarketShare, and also some of his successes. For example, despite the fact that data has grown significantly in recent years, business users still want faster results. This phenomenon led to efforts that supported larger amounts of data within his organization and demanded speed improvements - going from several minutes to sub-second responses. Satya will share some guiding principles that helped him successfully develop and deploy the systems his customers needed to be successful with their big data projects.

Using Qubole as the Data Lake for Programmatic Advertising

Mohave West

Tom Silverstrim, Senior Manager, Adobe Media Optimizer, Adobe Ad Cloud

Qubole has been the data warehouse of the DSP for the last six-plus years, and was selected as the ideal partner for mobilizing the considerable amount of diagnostic and base truth data contained within Amazon S3. From these origins, Qubole now powers our custom reporting infrastructure, machine learning algorithms, and user mapping reports, along with its evolving role in supporting system diagnoses and audits. We will touch on several use cases that demonstrate the flexibility and power of Qubole in democratizing data across the organization.

Highly Scalable and Flexible ETL Tool Built on Top of Cascading Framework

Aztec B

Navin Agarwal, Principal Engineer, BloomReach

At BloomReach we have 100+ e-commerce customers sharing product catalogs that range from a few megabytes to hundreds of gigabytes, which then need to be parsed and transformed. In this presentation we will talk about how we built a custom ETL transformation tool on top of a cascading framework that handles custom transformations and joins at scale and speed.

Building Data Functions at Poshmark: From KPI Monitoring to Enabling Social Graphs

Aztec C

Barkha Saxena, Vice President, Data & Analytics, Poshmark

Poshmark is the largest social marketplace for fashion in the U.S. where anyone can buy, sell, and share their personal style. With users engaging in 300M+ activities every day, data is a core asset at Poshmark. We began our data journey with very basic uses of data - monitoring high-level business KPIs. Four years later, we are now deploying data applications for actions such as enabling a balanced social graph and driving our homepage content based on real-time community activity. Join me in this session to hear key highlights from this incredible journey of building a data function at Poshmark, along with insights from the development of a people-matching algorithm and real-time user-driven homepage content.

Session IV 3:30 pm - 4:15 pm Location

Optimize for Reduced Big Data Partitioning Costs

Mohave East

Waad Aljaradt, Data Engineer, inMarket Media

Businesses that collect and process data can benefit greatly from partitioning their tables. Partitioning improves performance, increases query performance, and reduces the effort of rebuilding tables. Single partition queries can also be used to reduce the query load and avoid scanning the entire table.

However, transitioning large existing tables into partitioned tables can be cost-prohibitive. For example, we at inMarket media load billions of location records into multiple tables in our database to process these records through a pipeline of transformations. The resulting tables were not originally partitioned, and as time went on the decision not to partition the tables became increasingly expensive to maintain. We decided to partition our large existing tables to improve performance and reduce the costs of our queries. Initially, we thought that partitioning at this stage might be expensive. In my talk, I will give you a brief overview of the partitioning feature and explain the advantages and drawbacks of several different implementations of the partitioning process, as well as show how we were able reduce the cost of partitioning.

Data As Reliable As Running Water: Analytics and Machine Learning at Uber

Aztec A

Nikhil Joshi, Senior Product Manager, Data Infrastructure and Data Platforms, Uber

Every aspect of the Uber experience is powered by data - everything from in-app ETAs, menu recommendations, and map labeling to driver dispatch and customer support. In this talk, we will focus on the infrastructure and platforms that power data ingestion, storage, streaming/batch analytics, and machine learning for thousands of Operators, Data Scientists, and Engineers at Uber.

The 3S Method for Cluster Architecture Design

Mohave West

Justin Wainright, Systems Analyst, Oracle Data Cloud

This session highlights the model used within Oracle Data Cloud (ODC) for Apache Hadoop 2 and Apache Spark clusters. We'll talk about taking the guesswork out of cluster design, and about the keys for balancing cost and performance while minimizing administrative overhead.

An Interactive Discussion on Building a Data-Driven Culture

Aztec B

Wade Warren, Sr. Vice President, Global Engineering & TechOps, Wikia/FANDOM

Examples include:

1. From Verisign's "SiteFinder" debacle to the "Internet Threat Tracking Service."
2. Netflix: How we know more about what you love to watch than you do.
3. Wikia/Fandom: How to get you the most meaningful content and engaging experience.

Testing Spark Applications

Aztec C

Kurt Fehlhauer, Lead Database Architect, Activision

Apache Spark is a general-purpose computing engine for large-scale data processing. This talk is for data scientists and data engineers who want to understand the Apache Spark testing landscape within the Qubole environment. We will explore what it means to test an Apache Spark application, along with the tools and frameworks to make testing easier and more robust.

Session V — 4:30 pm - 5:15 pm — Location

Faster SQL on Apache Hadoop on Cloud Platforms for Ad-Hoc and Interactive Analysis

Mohave East

Rajat Venkatesh, Senior Director, Engineering, Qubole

Popular SQL on Hadoop engines like SparkSQL on Apache Spark, Hive, and Presto have become much faster on the cloud. This talk will explore the major features, architectural changes, and best practices to supercharge these SQL engines. The talk will also peek into the future for upcoming performance-related features.

Supercharging the Performance of Spark Applications

Aztec A

Venkat Sowrirajan, Software Engineer, Qubole

Apache Spark applications are difficult to tune for optimal performance, and the use of cloud stores like S3 as a truth-store makes things even more complex. This talk will briefly cover SparkLens (Spark tuning tool), Spark with Rubix (distributed cache), and direct-write for Hive tables and its performance numbers.

Using the Right Open Source Engine for the Right Job

Mohave West

Namit Jain, Senior Vice President, Engineering, Qubole

Qubole has built a service that runs multiple engines on multiple clouds. In this talk, we will discuss the journey and the lessons learned, including insights from usage across multiple engines. The session will cover a range of topics such as running a reliable service, optimizing performance, increasing mean time to failure (MTTF) through improved monitoring, and implementing practices for staggered rollouts and fallbacks.

Reimagining Data Platforms - Design Thinking and Innovation Workshop

Aztec C

Ankita Gautam, User Experience Designer, Qubole

Join us for a fun and interactive hands-on session to reimagine and design your ideal data platforms solution. Learn and use Design Thinking methodology and add this creative next-gen tool to your problem-solving arsenal! Become a part of this exclusive club of innovators. Limited seating.

Auto Tuning Twitter Hadoop Jobs or: Don't Touch That Analytics Dial! — Aztec B

Anton Panasenko, Software Engineer, Twitter & Ben Pence, Software Engineer, Twitter

Every day at Twitter, hundreds of thousands of Hadoop jobs transform and aggregate petabytes of data in our analytics stack. Historically, we've asked users to guess at tuning parameter values that affect how their Hadoop jobs run. For example, mapper and reducer counts, memory allocation, and intermediate serialization formats, among others. However, after looking at the values that users chose for tuning parameters in 2017, the data revealed that Hadoop jobs across our clusters were running sub-optimally and still not meeting users' Service Level Agreement (SLA) targets.

To address this problem, we implemented a service to automate the tuning of several of the most important Hadoop parameters, using historical per-job metrics to inform future runs. In this talk, we will review how the system works, some of the auto-tuning we've implemented so far, and what we have on our roadmap for the future.

6:30 – 9:30 pm	**Outdoor Dinner & Evening Event** You've spent the day learning how the wild west of big data is being won, now don your ten gallon hat and join us for a farm-fresh dinner and ice cold saloon drinks. Watch the sunset on the patio of a beautiful adobe building, and network with other leaders and innovators in the big data space. This is a party not to be missed.	Litchfield Lawn

Friday, April 13 — Location

7:30 - 8:45 am	**Breakfast**	Sachem
9:00 - 9:30 am	**Closing Keynote:** Kevin Kennedy, Chief Operating Officer, Qubole	Wigwam Ballroom

Session VI — 9:45 am - 10:30 am — Location

Kubernetes for Data Engineers — Mohave East

Rohit Agarwal, Software Engineer, Google

The talk will give an introduction to Kubernetes in general and then focus on topics relevant to data engineers. In particular, we will talk about how to run stateful workloads on Kubernetes and how to run machine learning workloads that use GPUs on Kubernetes

Big Data + Data Warehouse = Better Together — Aztec A

James Rowland-Jones, Principal Program Manager, Microsoft

Building a coherent platform for advanced analytics and reporting can feel quite overwhelming. A plethora of choices exist out there, and often it can feel like you are being asked to choose a side - it's almost like being asked to pick your favorite child! However, it doesn't have to be this way. Many of the services in a next-generation platform are actually complimentary, working together to deliver your next-generation analytical architecture.

In this session we will walk through the core components of a next-generation analytical platform architecture, discussing key decision points along the way. At the end you will have a clear and concrete understanding of how you can easily stand up an advanced analytical platform in minutes and bring demonstrable value to your users.

Enterprise Fabric – A Concept for the Essential Thread in Your Transformational Journeys — Mohave West

Dan Sutherland, Distinguished Engineer & CTO, Data Platforms, Global Business Services, IBM

This session will provide an overview of the enterprise fabric and the encapsulated view of the required capabilities. Some key components of the fabric include data and cognitive technologies. We will dive into the enterprise fabric-based architecture and why it is the core foundation for business transformation.

Where We're Going, We Don't Need Computers: End-to-End, Serverless Data Science — Aztec B

Alex Sadovsky, Senior Director of Data Science, Oracle Data & Cloud

Data scientists are expected to wear many hats in an organization. Many tasks often fall in the realm of data science - ingesting and cleaning data, managing data storage, creating scalable machine learning models, and publishing APIs to expose and schedule services for end users. This talk focuses on how to create end-to-end data science products that allow data scientists to focus on business logic, all while embracing nearly infinitely horizontally scalable data platforms.

To do this, we'll explore serverless cloud technologies at multiple levels of the data science pipeline such as serverless compute, workflow, containerized workloads, distributed on-demand machine learning, metrics tracking, and API as a service. At the end of this talk we'll have a prototype for an end-to-end machine learning system, on a scalable cloud platform, capable of processing petabytes of data and thousands of requests without the need for any freestanding servers.

Building Your Data Lake on AWS: Architecture and Best Practices — Aztec C

Paul Sears, Solutions Architect, AWS Partner Network, Amazon Web Services

As organizations aim to become more data-driven, data engineering teams must build architectures that can cater to the needs of diverse users - from developers and business analysts to data scientists. Each of these user groups employs different tools, has different data needs, and accesses data in different ways. Learn how to build and architect a data lake on AWS where different teams within your organization can publish and consume data in a self-service manner. Also learn about best practices for data curation, normalization, and analysis on Amazon object storage services.

Session VII — 10:45 am - 11:30 am — Location

Team Data Science Process (TDSP) and Azure Machine Learning — Mohave East

Erik Zwiefel, Advanced Analytics and AI Architect, Microsoft

TDSP is an agile data science process meant to keep data science and business teams working together. In this session, we'll explore the Team Data Science Process and walk through an example using Azure Machine Learning Services.

Email Text Classification: Building an End to End Data Product — Aztec A

Sasha Mushovic, Data Scientist, Return Path

Sasha will tell the story of building an end-to-end data product that feeds various parts of the Return Path business to optimize email programs for marketers. We will cover discovery, development, and production of an email classification model that uses Apache Spark to fit classifiers such as Random Forests and Support Vector Machines to read email text and classify the content. We will discuss the different methods of hyperparameter tuning and ensembling used, and will describe different stages of production from batch jobs in Qubole Scheduler and Apache Airflow to streaming in Apache Kafka. We will also reflect on what it means to be a full stack data scientist, and how data science teams can be empowered to own their own data products.

Democratizing the Data Pipeline — Mohave West

Zack Shapiro, Lead Data Architect, Nextdoor

Learn how the data team at Nextdoor.com stopped writing queries all day and developed a platform that empowered the entire company to build their own data pipelines.

Velocity Versus Volume — Aztec B

Sean Downes, Senior Data Scientist, Expedia

For technology companies, there is an inherent tension between streaming and batch processing. Real-time datastreams can transform a small input signal into an immediate response, but machine learning is most effective in batch. Modern data platforms can easily handle both streaming and batch jobs simultaneously. Balancing these two paradigms thus becomes a matter of design, and right now this interplay is thriving at the intersection of product and data science.

We discuss these dualities in the context of recommendations systems, some of our core products at Expedia. We'll sketch the design, architecture, tools, and metrics, as well as share our experience with our attempts at personalization. We'll merge the ideas behind multi-armed bandits and learning-to-rank to develop a novel recommendation system and give you the background needed to start building products in this rapidly evolving space.

The Dismal and Uncomfortable Science of Data Engineering: Building Out Big Data with Your Analytics Team — Aztec C

Charles Pritchard, Data Janitor, Jumis

While software services catch up, many big data projects rely on engineering resources, people with programming skills, and the culture around software development. Meanwhile, analysis teams are often oriented to serve customer engagement and finance managers, with an ethos and engagement style quite distinct from their counterparts in the school of computer science. As business and engineering sides of the house clash and cooperate, it's important to remember the human side of things, both in the data and in the delivery of insights. Let's talk about ways these capable business analysts and data engineers can work together, and ideals they can align toward.

11:30 am	**Departures**	
11:30 am	**Grab & Go Lunch**	Sachem