QCon SAN FRANCISCO by InfoQ

Architecture at QCon San Francisco 2026

Sessions

How Airbnb Guardrailed Its AI Customer Support Agent

Monday Nov 16

When Airbnb launched its AI customer support assistant, the feature made the news — the guardrails built to make it safe to ship did not. Yet an AI agent that talks to millions of customers, remembers a conversation across turns, and can trigger account actions is a

Weiping Peng

Distinguished Engineer @Airbnb, Previously Principal Architect @Salesforce for AI platform

---

Architecting Nubank's Global Financial Infrastructure

Monday Nov 16

Instant payment frameworks are transforming global finance, but few institutions have faced the infrastructure scale required by Brazil's Pix network.

Cat Swetel

Senior Director of Engineering @Nubank - General Manager for the Foundation Business Unit

---

Adapt or Drift: Resilience Engineering When AI Moves the Operating Point

Monday Nov 16

Resilient systems do not stay resilient by standing still. They survive by adapting. But adaptation comes with risk: under sustained pressure, organizations and systems can slowly drift toward failure while still appearing to operate normally.

Andrew Hatch

Engineering Leader and SRE Manager @Cisco ThousandEyes, With 25+ Years Building Software, Operations, SRE, and Platform Teams Across Australia, India, and the United States

---

QCon
SAN FRANCISCO by InfoQ

How to Build a Real-Time Voice Agent

Monday Nov 16

A voice agent looks like a chatbot with a microphone.

Sahil Yadav

Software Engineer @Together AI - Building Real-Time Production Voice Agents

---

Building Ultrafast LLM Inference at Scale: Disaggregated Inference across GPUs and Cerebras WSEs

Monday Nov 16

Ultrafast inference is a balancing act: users want tokens quickly, while providers need high aggregate throughput. The hardware choices that improve one can limit the other.

Nancy Zheng

VP of Engineering of AI Inference @Cerebras

Saturation: How Your Software Will Fail at Scale

Monday Nov 16

Even in the ethereal world of software, everything has a limit. And, once that limit is reached, very bad things can happen. In this talk, we will explore the problem of saturation: when a software system runs into one of these limits.

Lorin Hochstein

Staff Software Engineer @Airbnb, Writes @surfingcomplexity.blog, Previously @Netflix and Member of the Resilience in Software Foundation

---

Generation Got Cheap, Trust Didn't: Scaling Agentic Changes from Prompt to Prod at Roblox

Monday Nov 16

As AI agents write more of the code Roblox engineers push, the hard part is no longer generating changes; it's knowing which ones to trust.

Kushan Mehta

Principal Software Engineer @Roblox, Tech Lead for AI-Native Engineering

Jay Moorthi

Senior Engineering Manager, Diagnostics @Roblox

Breaking Down the Walls to Python & Big Data + AI Performance with Transpilation and LLMs

Tuesday Nov 17

Many of us have a love hate relationship with Apache Spark, especially when it comes to PySpark. Especially in Python, the UDF performance can be a dumpster fire.

Holden Karau

Data Engineer @Snowflake & Co-Founder of Fight Health Insurance, Previously @Netfilx, @Apple, and @Google

---

Inside Asimov: BlackRock's Platform Architecture for Applied AI at Scale

Tuesday Nov 17

BlackRock's Portfolio Management Group is transforming how investment teams conduct research by combining deep investment expertise with shared AI capabilities.

Dan Wolf

Head of Platform Engineering for PMG Tech @BlackRock

Sri Annapurna Bhupatiraju

Director and Principal Engineer @BlackRock, One of the Founding Members of BlackRock AI Labs

---

One Does Not Simply Add Synchronous Writes - Bridging Batch and Real Time with Apache Iceberg

Tuesday Nov 17

What happens when a batch ingestion platform needs to behave like an operational database? Datadog's Reference Tables product was designed for asynchronous uploads of large datasets, but customers increasingly needed synchronous row-level edits, without sacrificing batch scale.

Paige Elinson
Staff Software Engineer @Datadog

Autoscaling 800 Valkey Clusters: Distributed Systems Lessons from Scaling Stateful Fleets

Tuesday Nov 17

Autoscaling stateless services is a solved problem. In a sharded datastore, a resize is not a capacity change — it's moving ownership of key ranges between nodes while the cluster serves live traffic, and rebalancing takes long enough that reacting to a traffic peak is already too late.

Kishor Yadav Kommanaboina
Staff Software Engineer @Snapchat - Leading Caching and Storage Infrastructure

---

Building a StreamHouse: Turning Kafka Streams into a Millisecond Query Layer

Tuesday Nov 17

AI agents have compressed business decisions from minutes to seconds. Data systems need to keep up.

RJ Atwal
Internal Database Engineer, Principal Engineer @Confluent, Previously Founding Engineer @MotherDuck and @AWS

When Your Users Are Agents: Lessons from a Distributed Postgres Platform

Tuesday Nov 17

Distributed systems are built around assumptions about workload behavior: connections have reasonable lifetimes, retries eventually stop, traffic spikes have recognizable causes, and application code produces somewhat predictable query patterns.

Gwen Shapira
Co-Founder and CPO @Nile, Previously Engineering Leader @Confluent, PMC Member @Kafka, &

Committer Apache Sqoop

Iceberg V4: The Next Chapter in Table Formats

Tuesday Nov 17

Apache Iceberg V3 adoption is growing and the community is already shaping what comes next. This talk walks through the emerging scope and priorities for the V4 table spec and progress of the proposals currently underway.

Daniel Weeks

MTS @Databricks, Co-Founder of Tabular and Co-Creator of Apache Iceberg, 20+ Years Designing & Developing Large-Scale Distributed Systems

---

Adaptive Systems in Production: What Recommendation Systems Can Teach Us About Agents

Tuesday Nov 17

As organizations race to build AI agents, many teams are encountering challenges that feel new: evaluation uncertainty, feedback loops, behavioral drift, exploration versus exploitation, and maintaining user trust in systems that continuously adapt.

But these challenges are not new.

Mallika Rao

Senior Engineering Manager @Zocdoc, Previously @Netflix, @Twitter and @Walmart

---

How Netflix Retrofitted Its SDLC for AI Coding Agents

Wednesday Nov 18

Agentic tools amplify what's already in the codebase. The agent works through your code, your architecture decisions, and the documentation around them. It follows what it can read and guesses at the rest.

Mark Khuzam

Senior Software Engineer @Netflix, Previously Led the Consumer Web Platform Team @OpenTable

QCon
SAN FRANCISCO by InfoQ

No More Small Systems, How AI May Bring Large System Problems to Everyone

Wednesday Nov 18

Once you figure out how to use AI to make your organization productive, you will soon find you have a whole new set of problems. While some of those problems will be novel thanks to LLMs, many of them are just the result of there being "more stuff".

Adam Bender

Technical Lead for Enterprise Engineering @Google

How LinkedIn Runs Compute: Self-Managed Kubernetes on 500,000+ Bare-Metal Servers

Wednesday Nov 18

LinkedIn's Kubernetes-based compute platform spans more than 500k bare-metal machines, 5M pods, with thousands of developers doing more than 110k deploys a week. We don't expose raw Kubernetes to developers and instead, provide a curated set of platform offerings.

Ronak Nathani

Principal Staff Software Engineer, Compute Infra @LinkedIn, Podcast Host @Software Misadventures

---

Avoiding Prompt Debt & Other Context Failures

Wednesday Nov 18

As agents evolve, prompts get unruly. Edge cases are patched with one-line hotfixes, or with stacked instructions of escalating severity when an issue proves stubborn. Over time, those instructions pile up into a brittle artifact, overfit to an aging model.

Drew Breunig

CEO & Cofounder CMPND, Editor of O'Reilly's "Context Management Handbook", Writer on AI & Data at dbreunig.com, Previously Data Science & Product Lead PlaceIQ
