



Big Data in Social Media

Exploring TikTok's Data-Driven
Success Story

Ngoc My Nguyen

June 2025

TikTok

Business Domain: Media & Entertainment



Social Media Platform

Enables users to create, share, and discover short-form videos, typically 15 to 60 seconds long => connects people around the world



E-Commerce & Digital Marketing

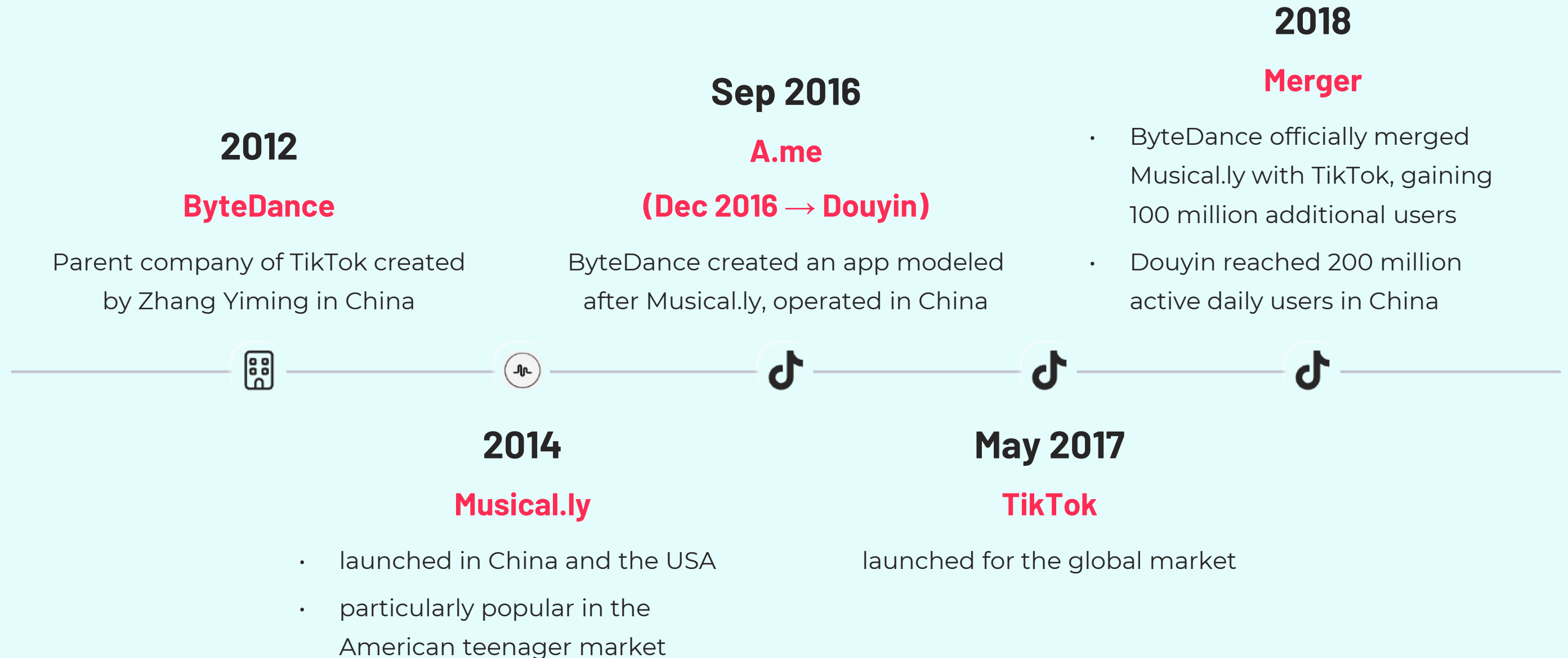
Enables creators to sell products through live streaming and allows brands to reach consumers directly



AI-driven Technology

Powered by a sophisticated recommendation algorithm that personalizes content delivery through the "For You" feed

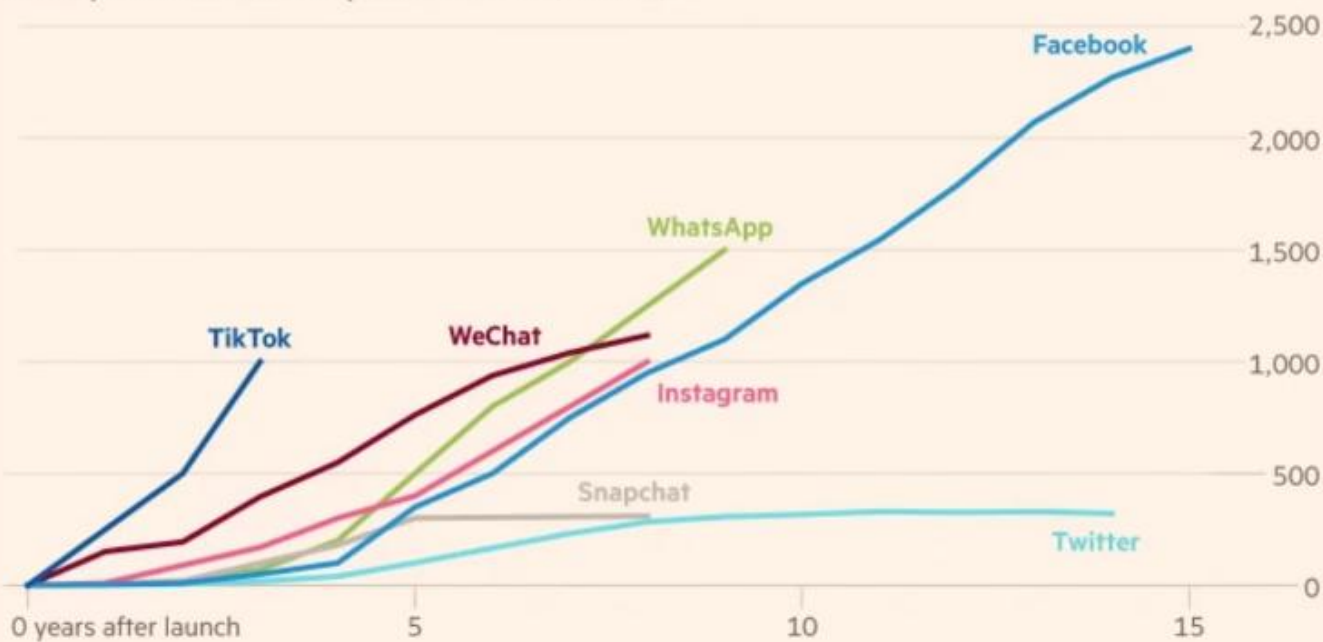
The History of TikTok



TikTok's Rapid Rise

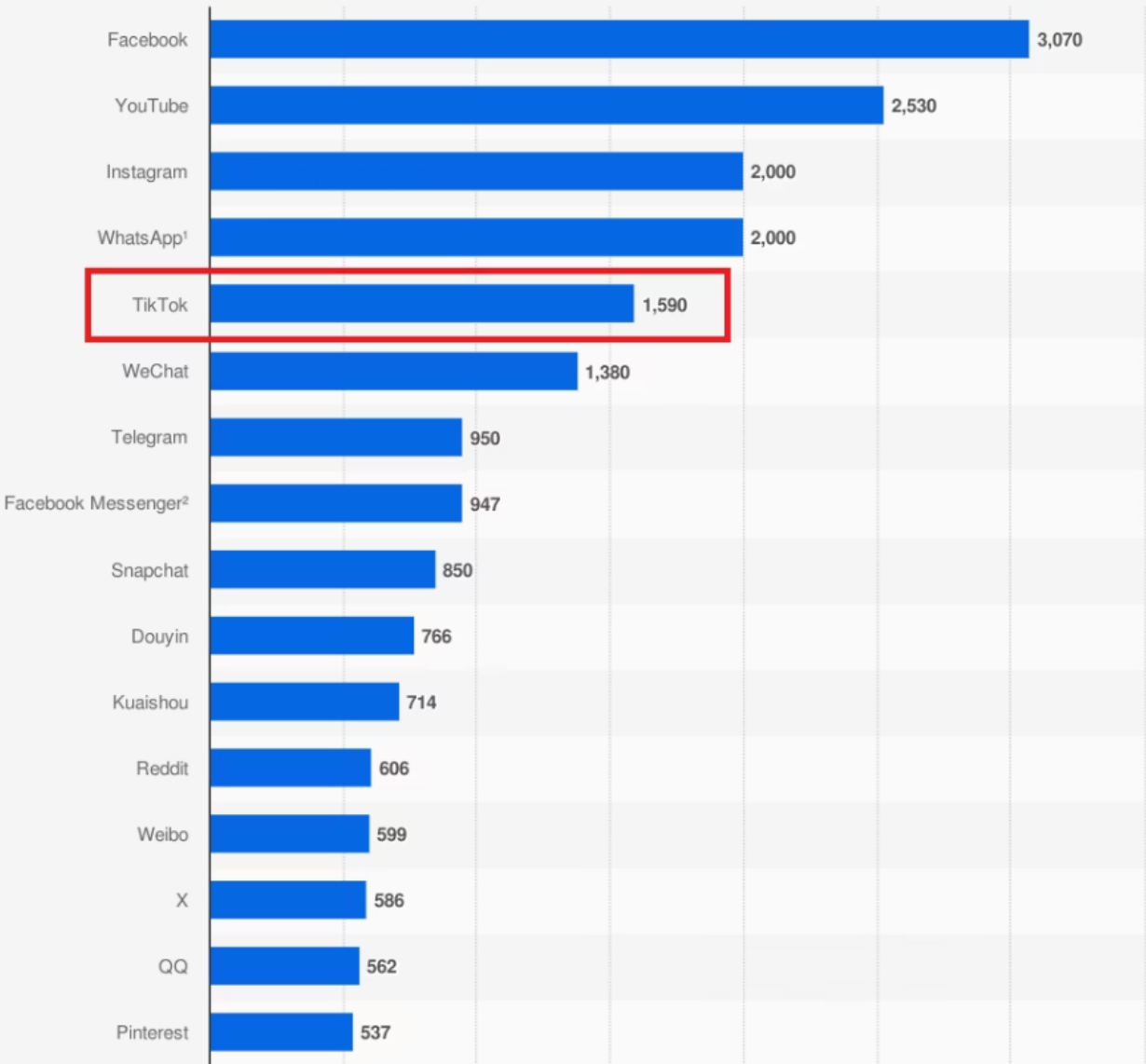
TikTok has reached 1bn users faster than any other social media app

Monthly active users since product launch (millions)



Source: FT research
© FT

Most popular social networks worldwide as of February 2025, by number of monthly active users (in millions)



Sources
We Are Social; DataReportal; Meltwater
© Statista 2025

Additional Information:
Worldwide; DataReportal; February 2025; social networks and messenger/chat app/voip included; figures for TikTok does not include Douyin

Big Data Use Cases



Personalized Recommendations

AI-driven content matching to user preferences



Content Moderation

Automated detection of policy violations



Ad Optimization

Precision targeting for advertiser campaigns



Trend Analysis

Real-time detection of viral content



Primary Data Sources

TikTok's algorithms process billions of data points daily to create personalized experiences.



Personal Data

e.g. account information, messages, contacts & connections, purchase information



User Behavior

e.g. engagement data (comment, like, share, subscribe, swipe, pause), scroll speed, watch time, interaction frequency, search history



Media Content

Videos, audio, hashtags, captions, and creator metadata.



Contextual Data

Time stamps, geographic data, device specifications, and network conditions for optimized delivery.



Social Graph

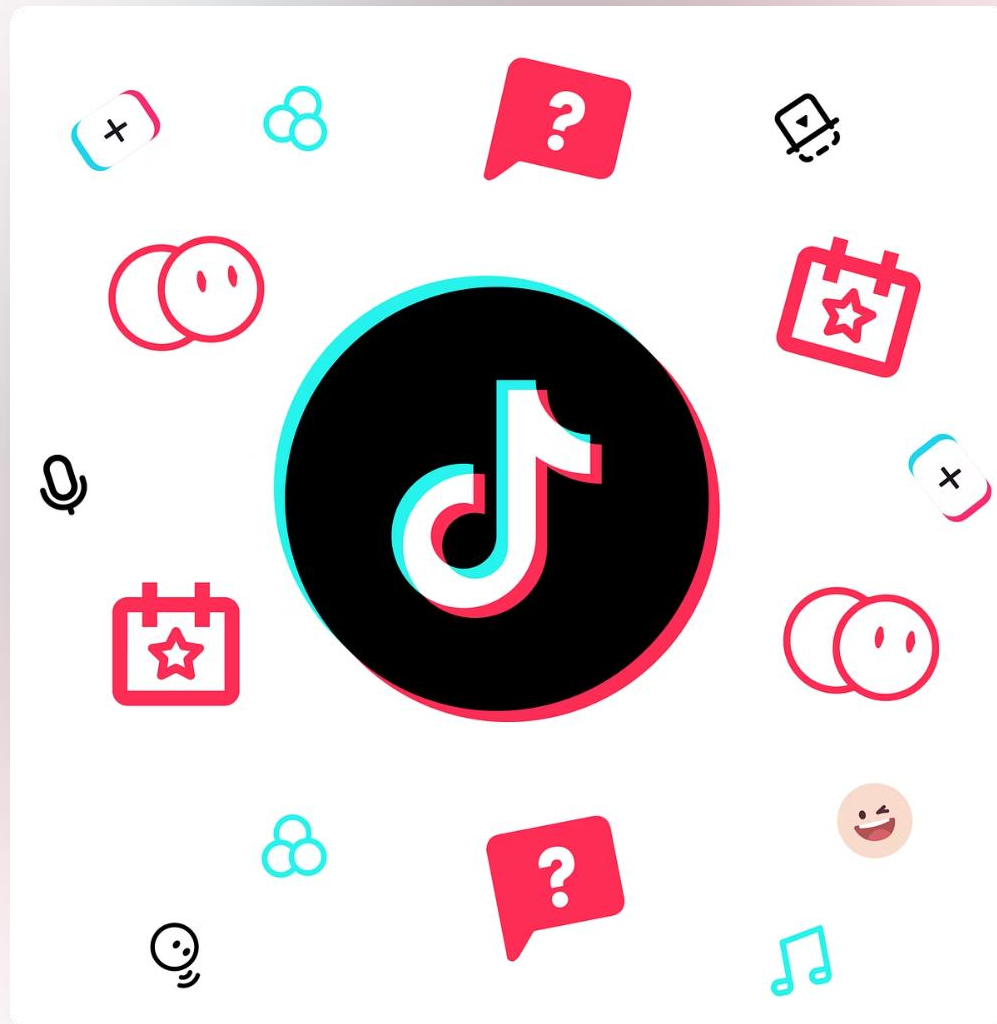
Follow relationships, shared content, and interaction networks.



Different Data Types


Sensitive - Structured



Semi-structured - Unstructured




Privacy & Data Security

TikTok's data is hosted in multiple locations around the world



User Location 	Default Storage Location 
United States	United States - Oracle Cloud Infrastructure
Europe	Dublin (2023) & Norway (anticipated 2024)
Rest-of-World	U.S. (Virginia), Singapore, and Malaysia



Source: [TikTok Facts: How we secure personal information and store data - Newsroom | TikTok](#)

TikTok is facing increasing limits and bans on a global scale



TikTok has been officially banned across the United States since **January 19, 2025**, due to the US government's concerns over **potential collection and influence operations** by a **government** of the People's Republic of **China**.

Delayed

[Learn more about countries that have banned TikTok](#)

IT Architecture (High Level)

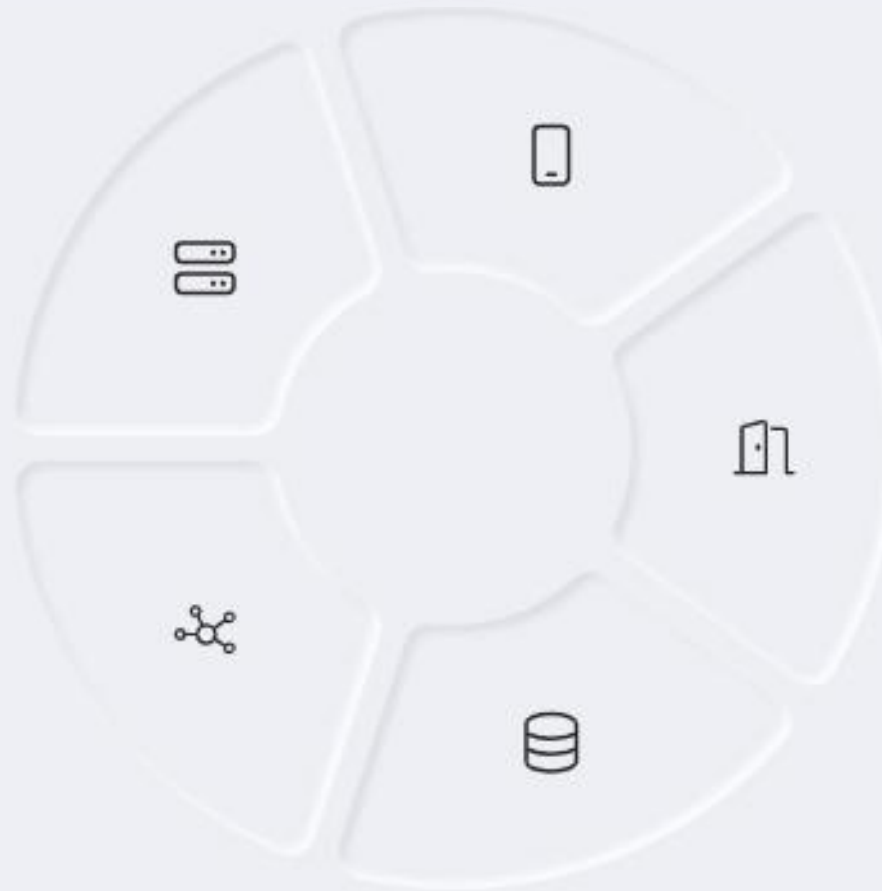
TikTok operates as a sophisticated distributed system with a multi-layered architecture designed for maximum efficiency.

API Gateway Layer

Request routing and authentication

Content Delivery Network (CDN)

Global video distribution



Client Layer

Mobile apps and web interfaces that users interact with directly

Application Layer

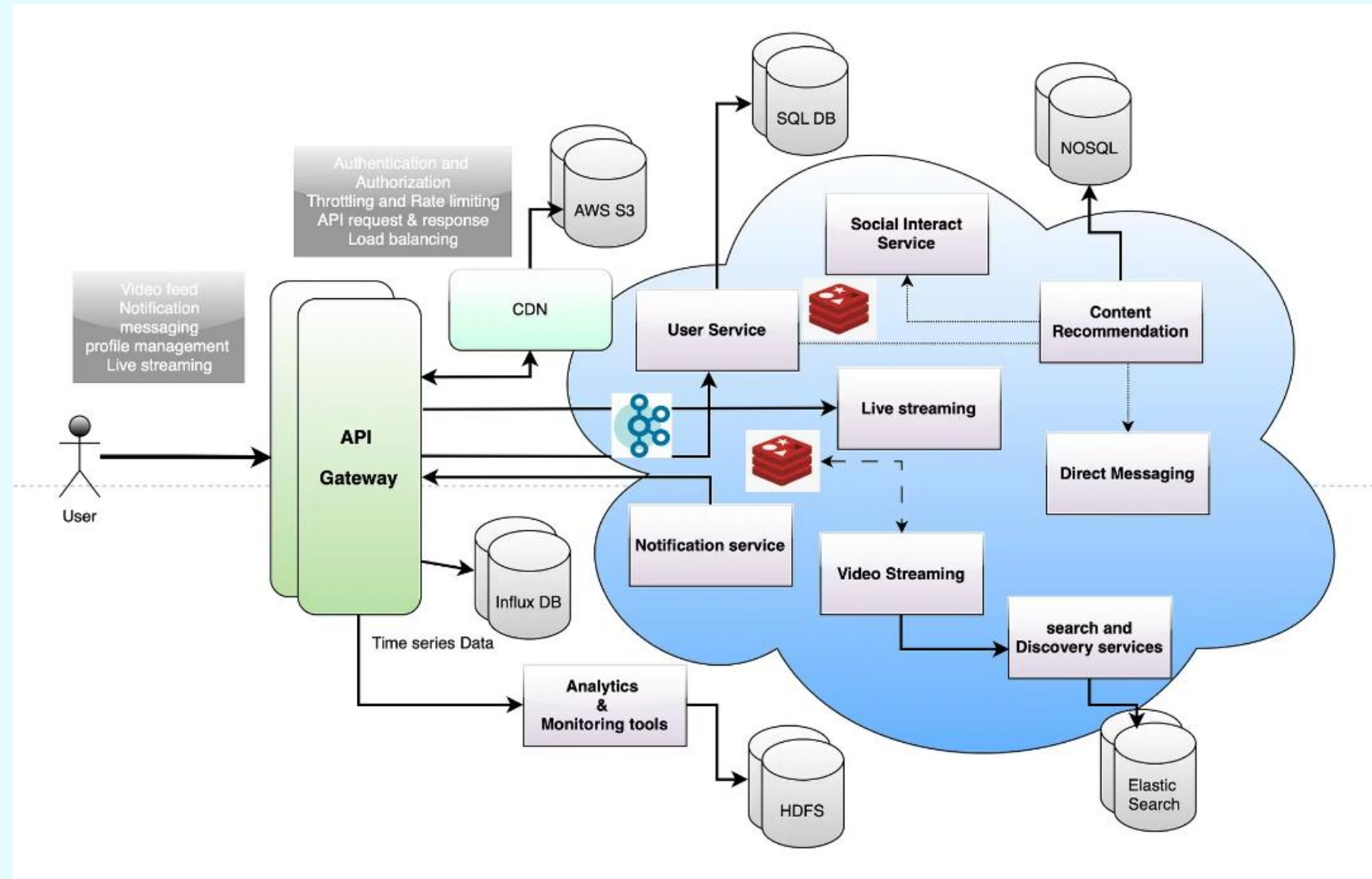
Independent microservices handling specific functionality

Data Storage Layer

Distributed storage and processing systems

⚠ Very few official documentations about their IT architecture, only articles from individual authors

High Level Design of TikTok



Source: [Design Of TikTok. TikTok is a social media platform for... | by Santosh P. | Medium](#)

How to verify the information?

https://lifeattiktok.com/search/7278873857032128805

Big data

Big Data Ecommerce Recommendation Infrastructure

Location: Singapore
Employment Type: Full-time
Job Code: A1319

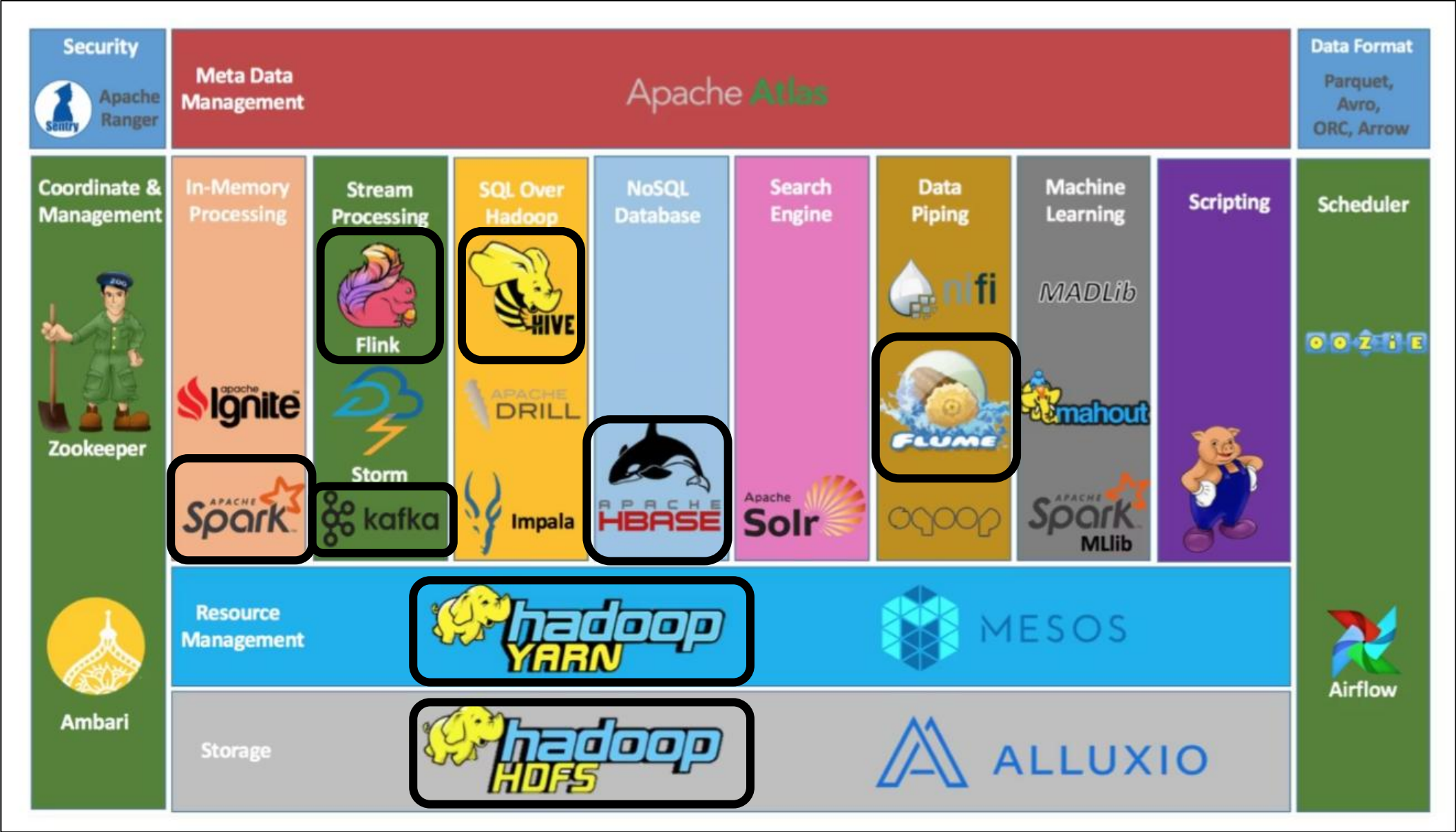
Apply to this

Qualifications

- Bachelor's degree in Computer Science or related field
- Familiar with distributed systems like Hadoop, HBase, RocksDB, etc.
- Familiar with SQL and NoSQL databases
- Strong coding and troubleshooting skills
- At least 5 years of relevant experience
- Deep understanding of streaming computing systems, with formal production experience in developing TB-level real-time computing systems. Proficient in modules like FlinkDataStream, FlinkSQL, FlinkCheckpoint, FlinkState, and preferably with experience in reading Flink source code.
- Experience in data lake development is preferred. Familiar with at least one data lake technology such as Hudi, Iceberg, DeltaLake, and preferably with experience in reading their source code.
- Willingness to tackle problems without clear answers, with a strong passion for learning new technologies.
- Experience in handling PB-level data is a plus.
- Familiarity with other big data systems is preferred, including YARN, K8S, Spark, SparkSQL, Kudu, and others.
- Experience in storage systems such as Hbase, Cassandra, RocksDB.

TikTok is committed to creating an inclusive space where employees are valued for their skills, experiences, and unique perspectives. Our platform connects people from across the globe and so does our workplace. At TikTok, our mission is to inspire creativity and bring joy. To achieve that goal, we are committed to celebrating our diverse voices and to creating an environment that reflects the many communities we reach. We are passionate about this and hope you are too.

TikTok's operational backbone: Hadoop Ecosystem



Other Big Data Technologies in Use

Key-Value Database



Columnar Database



DataLake

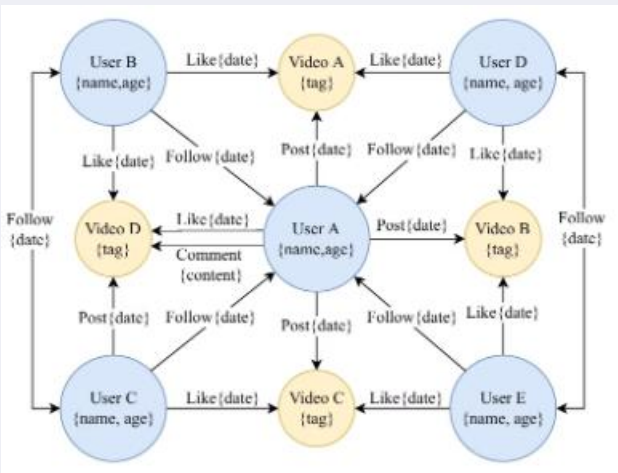


Search Engine



A distributed search and analytics engine based on Lucene, used for **full-text search**, **log analysis**, and more.

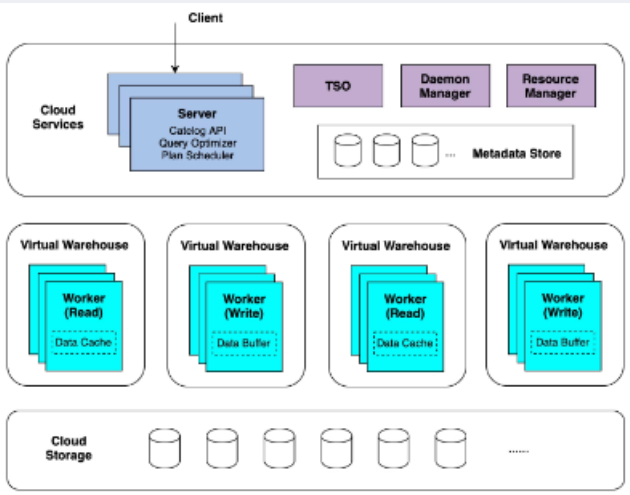
Graph Database



ByteGraph (proprietary)

[ByteGraph: a high-performance distributed graph database in ByteDance | Request PDF](#)

Data Warehouse



ByConity (open-sourced)

real-time data warehouse, based on ClickHouse but cloud-native.

ByConity is available on GitHub

Artificial Intelligence & Analytics



Deep Learning

Powers the recommendation engine, e.g. **Monolith**



Natural Language Processing

Analyzes captions and comments

Monolith is available on [GitHub](#)



Computer Vision

Enables content moderation and scene recognition

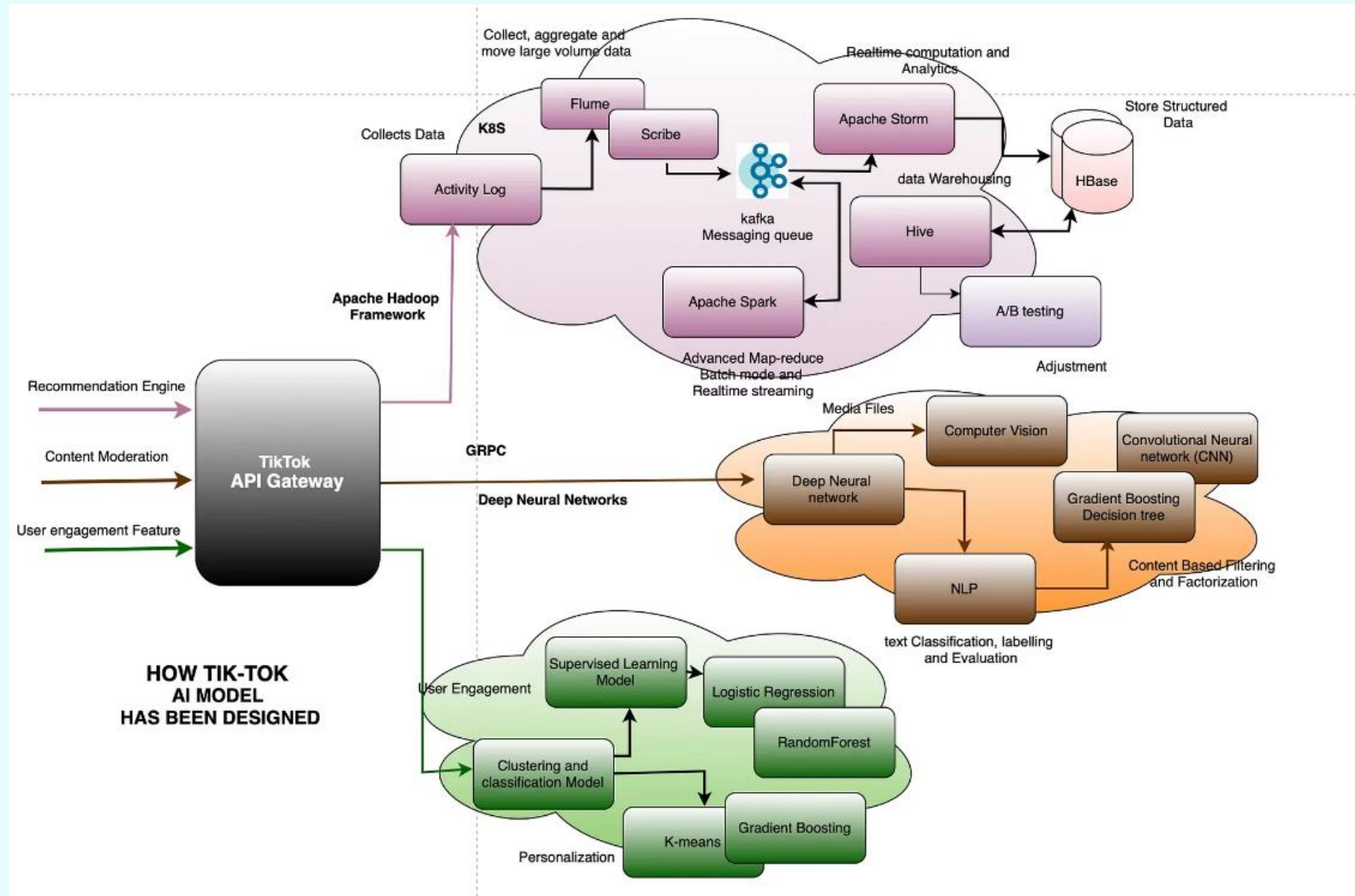


Predictive Analytics

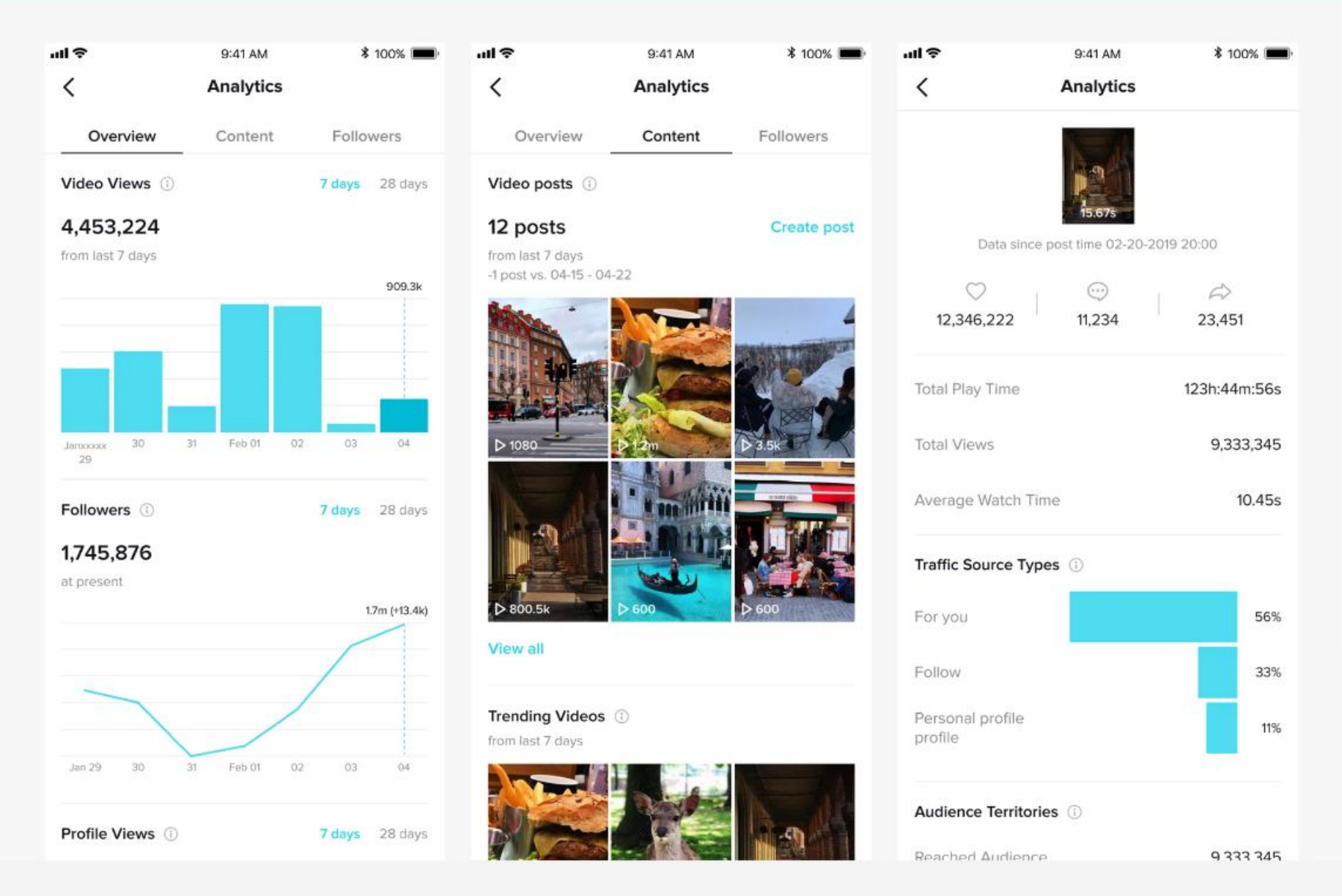
Forecasts engagement and trends

i With the goal to deliver an addictively smooth user experience to over a billion users worldwide, while still ensuring security and compliance.

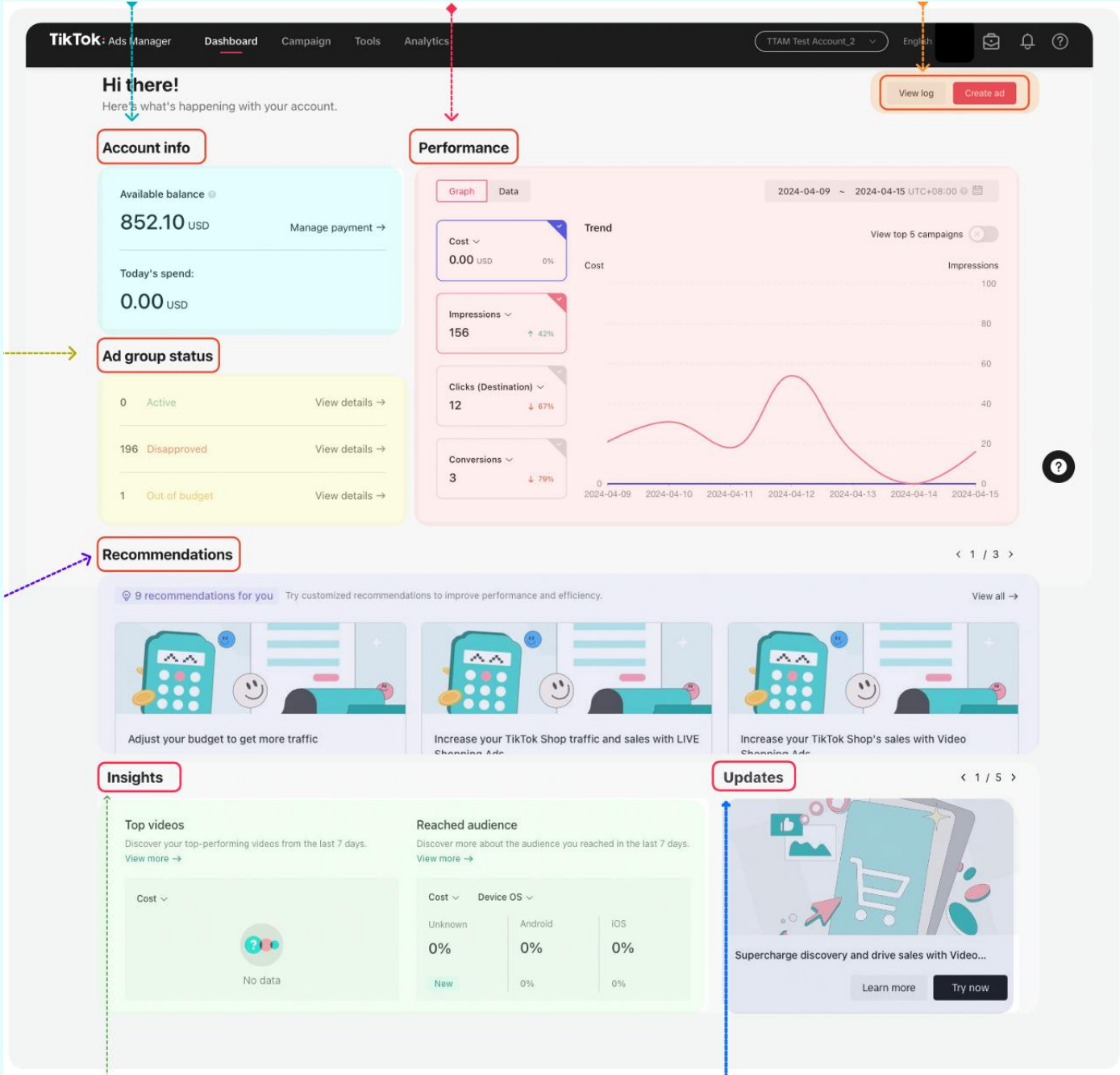
Summary: Big Data & AI in TikTok



Data Visualization & Dashboards for Creators



Data Visualization & Dashboards for Business



Source: [TikTok Ads Manager User Playbook](#)



Thank you!

Any questions?