LinkedIn: Social Engineering with Big Data Applications

Hannes Bähr

Big Data Technologies

02.06.2025

Content

1. LinkedIn: Platform & Big Data Relevance
2. Data Sources & Data Types
3. IT Infrastructure & Big Data Technologies
4. Data Analytics & AI
5. Dashboards & Visualization
6. Big Data Use Cases at LinkedIn
7. Summary

1. LinkedIn: Platform & Big Data Relevance

About LinkedIn

- World's largest professional networking platform: 950+ million users
- Connects professionals, recruiters, companies, and job seekers

Business Sector

- Operates in social networking, recruitment, and B2B advertising
- Revenue from subscriptions (Premium), talent solutions, and ads

Why Big Data is Central

- User interactions, connections, profile updates, job searches generate large data volume
- Big data powers core features like:
    - People You May Know
    - Job Recommendations
    - Feed Ranking & Content Personalization

2. Data Sources & Data Types

Data Sources

- User Profiles: Information such as names, job titles, education, skills, and locations
- User Interactions: Data from likes, comments, shares, messages, and connection requests
- Job Postings: Details about job titles, descriptions, company information, requirements…
- Company Pages: Information on company size, industry, posts, and employee data
- Activity Logs: Records of user sessions, clickstreams, and search queries
- Connections Graph: Data representing the network of user connections

Data Types

- Structured Data: Tabular data like user profiles and job listings
- Semi-Structured Data: JSON/XML formats from logs and messages
- Unstructured Data: Text from posts, messages, and comments
- Graph Data: Network data representing user connections and interactions
- Time-Series Data: Chronological data from user activities and interactions

3. IT Infrastructure & *Big Data Technologies* (1)

Batch Processing

Batch Processing Engines

- *Apache Spark* performs distributed data processing and transformation
- *Apache Hive* enables SQL-based queries on large datasets in Hadoop

Data Storage

- *HDFS* stores petabyte-scale structured and semi-structured data

Workflow Orchestration

- *Azkaban* manages ETL workflows and batch job scheduling

Metrics and Monitoring

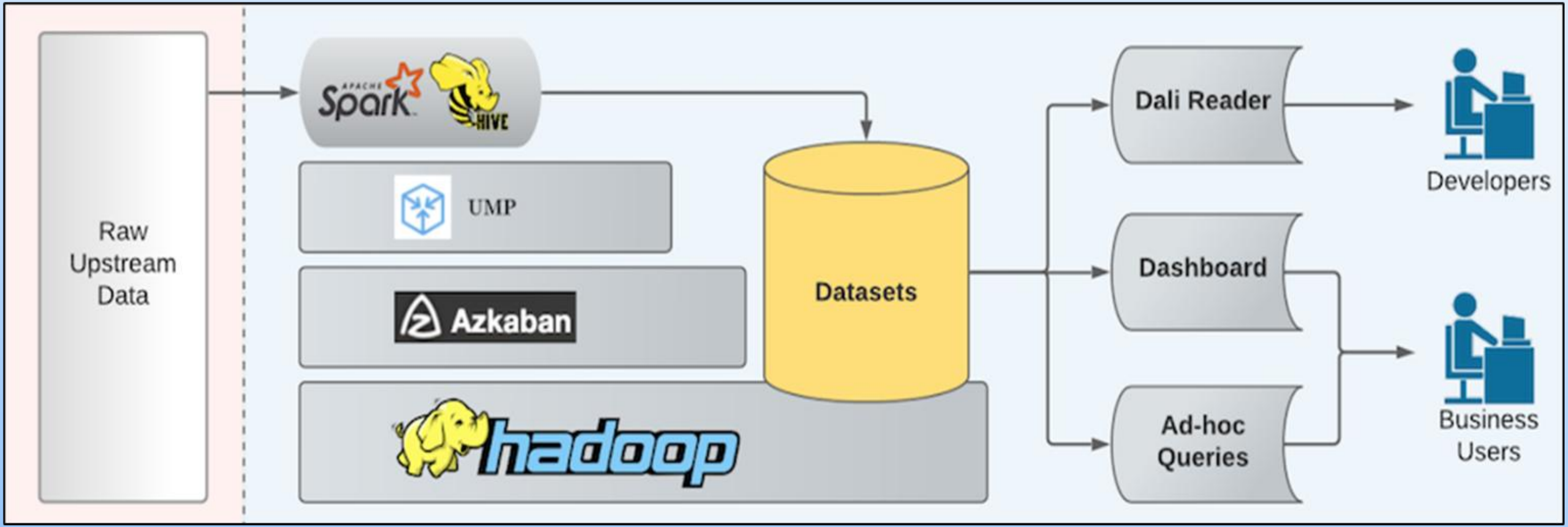
- *UMP* tracks pipeline health and operational metrics

Developer Access

- *Dali Reader* allows engineers to query and explore Hadoop datasets efficiently

3. IT Infrastructure & *Big Data Technologies* (1)

Batch Data Processing Flow



3. IT Infrastructure & *Big Data Technologies* (2)

Real-Time Processing

Real-Time Ingestion

- *Apache Kafka* ingests millions of real-time events (clicks, interactions, views) with low latency

Change Data Capture (CDC)

- *Brooklin* captures and streams changes from traditional databases like MySQL and other

Stream Processing

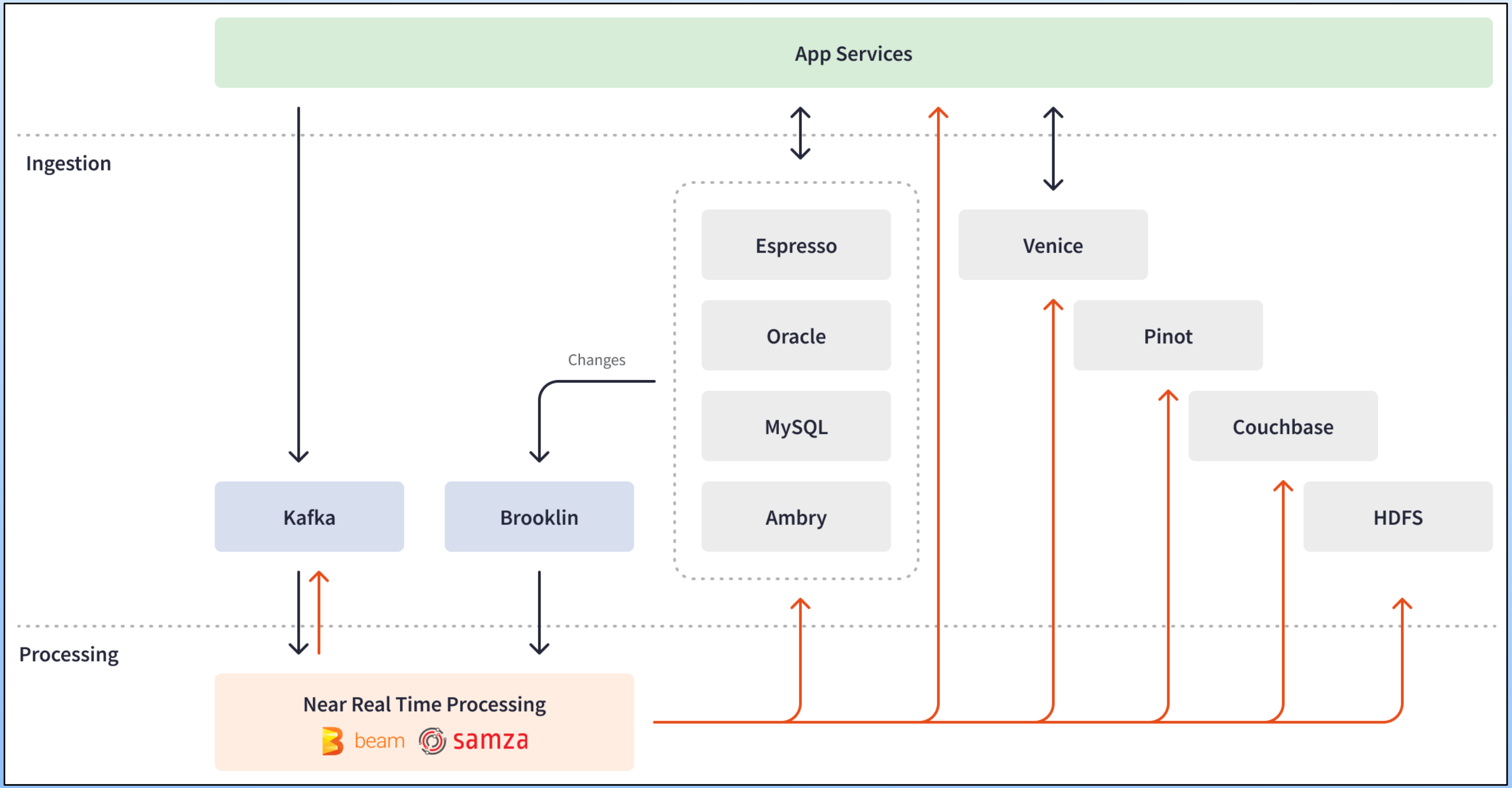
- *Apache Samza* processes real-time data from and to Kafka
- *Apache Beam* provides a unified programming model across batch and stream

Real-Time Analytics and Serving

- *Venice* and *Couchbase* serve online features to applications
- *Apache Pinot* supports low-latency OLAP queries for dashboards
- *HDFS* stores persistent data copies

3. IT Infrastructure & Big Data Technologies (2)

Real-Time Processing Flow



4. Data Analytics & AI

Machine Learning Applications

- *People You May Know*: Uses *LiGNN Framework* (graph-based system) to recommend new connections based on user network patterns and interactions
- *Skill Endorsements*: Recommends relevant skills using behavioral and network-based predictions
- *Job Recommendations*: Matches users to job listings based on profile, behavior, and similarity metrics

Forecasting and Time Series Analysis

- *Greykite*: LinkedIn’s open-source Python library for time series forecasting used in business planning

Natural Language Processing

- *Deep Job Understanding*: DL models analyze job postings to improve search and matching accuracy

AI-Powered Products

- *AI Hiring Assistant*: Automates job description writing and candidate matching for recruiters, powered by ML and NLP models

Real-Time Analytics and Processing

- *Apache Pinot*: Powers internal dashboards, real-time analytics, A/B test monitoring, and ML model feedback loops with sub-second query latency

5. Dashboards & Visualization

Internal Dashboards & Tools (for data scientists, analysts, engineers, …)

- *Pinot Dashboards*: Real-time analytics dashboards powered by Apache Pinot, used to monitor metrics, A/B test performance, and model behavior
- *UMP*: Visualizes system-level and business metrics across pipelines
- *WhereHows*: Metadata explorer for data lineage, table dependencies, and schema tracking
- *Third-Party Tools*: Tableau and Superset are used for custom BI dashboards

User-Facing Visualizations (for member, recruiters, and companies)

- Profile Analytics: Users see who viewed their profile and how often they appear in search
- Job Ad & Content Performance: Recruiters and marketers get dashboards showing impressions, clicks, and engagement
- Skills Insights & Career Explorer: Visual tools help users understand in-demand skills and career transitions based on LinkedIn data

6. Big Data Use Cases at LinkedIn

People You May Know

- One of LinkedIn's flagship features
- Powered by large-scale graph processing by using LiGNN
- Consumes real-time and batch data from connections, interactions, and profiles

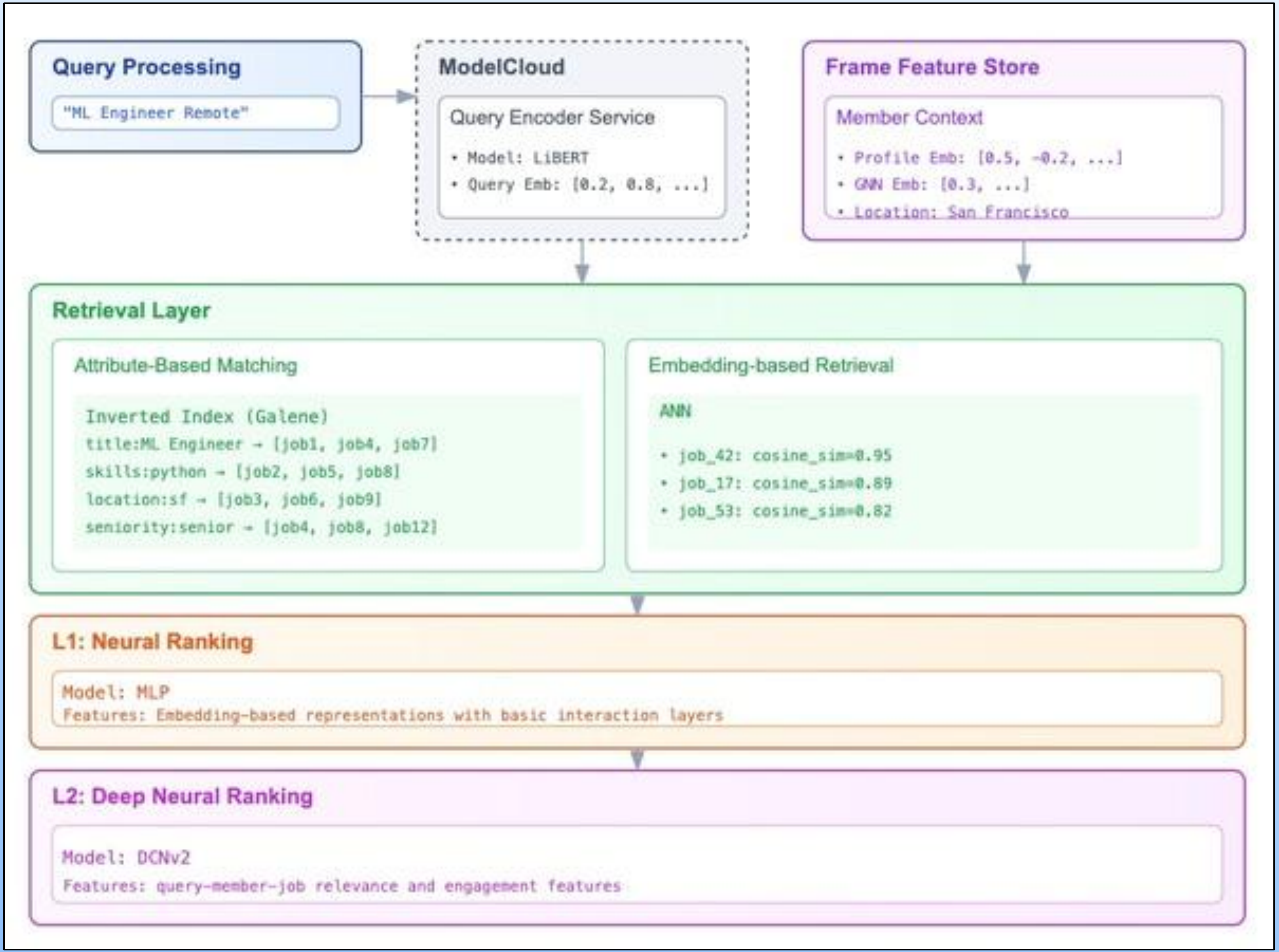
Job Recommendation Engine

- Uses a combination of NLP (parse job descriptions), filtering, and user behavior data
- Backed by real-time processing (Kafka, Samza) and batch retraining pipelines (Spark)
- Includes Deep Job Understanding system for semantic matching

AI Hiring Assistant

- AI-based tool for recruiters, automates job description generation and candidate matching
- Integrates ML models with user-facing dashboards and real-time data feedback

6. Big Data Use Cases at LinkedIn



Query Processing

"ML Engineer Remote"

ModelCloud

Query Encoder Service

- Model: LiBERT
- Query Emb: [0.2, 0.8, ...]

Frame Feature Store

Member Context

- Profile Emb: [0.5, -0.2, ...]
- GNN Emb: [0.3, ...]
- Location: San Francisco

Retrieval Layer

Attribute-Based Matching

Inverted Index (Galene)
title:ML Engineer → [job1, job4, job7]
skills:python → [job2, job5, job8]
location:sf → [job3, job6, job9]
seniority:senior → [job4, job8, job12]

Embedding-based Retrieval

ANN

- job_42: cosine_sim=0.95
- job_17: cosine_sim=0.89
- job_53: cosine_sim=0.82

L1: Neural Ranking

Model: MLP
Features: Embedding-based representations with basic interaction layers

L2: Deep Neural Ranking

Model: DCNv2
Features: query-member-job relevance and engagement features

7. Summary

LinkedIn's Platform

- A leading professional network using big data to personalize user experiences and power business tools

Data Architecture

- A hybrid system combining batch and real-time processing

Big Data Technologies

- Tools like Gobblin, Pinot, Azkaban, Helix, and Beam orchestrate and analyze data at scale

AI & Analytics

- ML and NLP models drive features like job recommendations, People You May Know, and AI hiring assistants

Dashboards & Visualizations

- Real-time analytics and user-facing dashboards support both internal decision-making and external engagement

Thanks for your attention!