Accelerating autocomplete suggestions at Salesforce Search

Georgios Balikas, Guillaume Kempf, Marc Brette

Twitter: @gbalikas

Forward-Looking Statements

This presentation contains forward-looking statements about, among other things, trend analyses and future events, future financial performance, anticipated growth, industry prospects, environmental, social and governance goals, and the anticipated benefits of acquired companies. The achievement or success of the matters covered by such forward-looking statements involves risks, uncertainties and assumptions. If any such risks or uncertainties materialize or if any of the assumptions prove incorrect, Salesforce's results could differ materially from the results expressed or implied by these forward-looking statements. The risks and uncertainties referred to above include those factors discussed in Salesforce's reports filed from time to time with the Securities and Exchange Commission, including, but not limited to: the impact of, and actions we may take in response to, the COVID-19 pandemic, related public health measures and resulting economic downturn and market volatility; our ability to maintain security levels and service performance meeting the expectations of our customers, and the resources and costs required to avoid unanticipated downtime and prevent, detect and remediate performance degradation and security breaches; the expenses associated with our data centers and third-party infrastructure providers; our ability to secure additional data center capacity; our reliance on third-party hardware, software and platform providers; the effect of evolving domestic and foreign government regulations, including those related to the provision of services on the Internet, those related to accessing the Internet, and those addressing data privacy, cross-border data transfers and import and export controls; current and potential litigation involving us or our industry, including litigation involving acquired entities such as Tableau Software, Inc. and Slack Technologies, Inc., and the resolution or settlement thereof; regulatory developments and regulatory investigations involving us or affecting our industry; our ability to successfully introduce new services and product features, including any efforts to expand our services; the success of our strategy of acquiring or making investments in complementary businesses, joint ventures, services, technologies and intellectual property rights; our ability to complete, on a timely basis or at all, announced transactions; our ability to realize the benefits from acquisitions, strategic partnerships, joint ventures and investments, including our July 2021 acquisition of Slack Technologies, Inc., and successfully integrate acquired businesses and technologies; our ability to compete in the markets in which we participate; the success of our business strategy and our plan to build our business, including our strategy to be a leading provider of enterprise cloud computing applications and platforms; our ability to execute our business plans; our ability to continue to grow unearned revenue and remaining performance obligation; the pace of change and innovation in enterprise cloud computing services; the seasonal nature of our sales cycles; our ability to limit customer attrition and costs related to those efforts; the success of our international expansion strategy; the demands on our personnel and infrastructure resulting from significant growth in our customer base and operations, including as a result of acquisitions; our ability to preserve our workplace culture, including as a result of our decisions regarding our current and future office environments or work-from-home policies; our dependency on the development and maintenance of the infrastructure of the Internet; our real estate and office facilities strategy and related costs and uncertainties; fluctuations in, and our ability to predict, our operating results and cash flows; the variability in our results arising from the accounting for term license revenue products; the performance and fair value of our investments in complementary businesses through our strategic investment portfolio; the impact of future gains or losses from our strategic investment portfolio, including gains or losses from overall market conditions that may affect the publicly traded companies within our strategic investment portfolio; our ability to protect our intellectual property rights; our ability to develop our brands; the impact of foreign currency exchange rate and interest rate fluctuations on our results; the valuation of our deferred tax assets and the release of related valuation allowances; the potential availability of additional tax assets in the future; the impact of new accounting pronouncements and tax laws; uncertainties affecting our ability to estimate our tax rate; uncertainties regarding our tax obligations in connection with potential jurisdictional transfers of intellectual property, including the tax rate, the timing of the transfer and the value of such transferred intellectual property; uncertainties regarding the effect of general economic and market conditions; the impact of geopolitical events; uncertainties regarding the impact of expensing stock options and other equity awards; the sufficiency of our capital resources; our ability to comply with our debt covenants and lease obligations; and the impact of climate change, natural disasters and actual or threatened public health emergencies, including the ongoing COVID-19 pandemic.

092821

Outline

- Autocomplete
- Salesforce Search backend
- Implemented Solution with DL
- Results and closing remarks

Given a query prefix, autocomplete suggests possible extensions of the prefix.
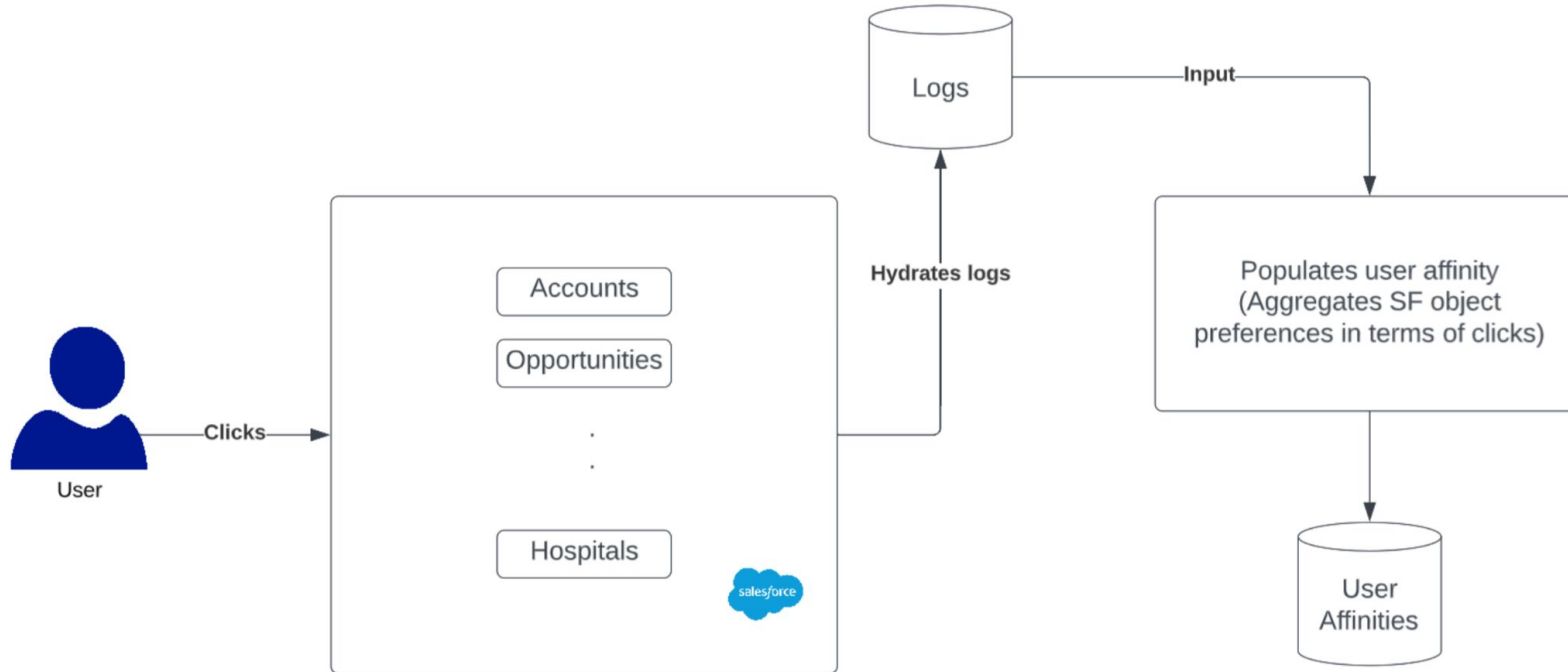
An abstract architecture diagram
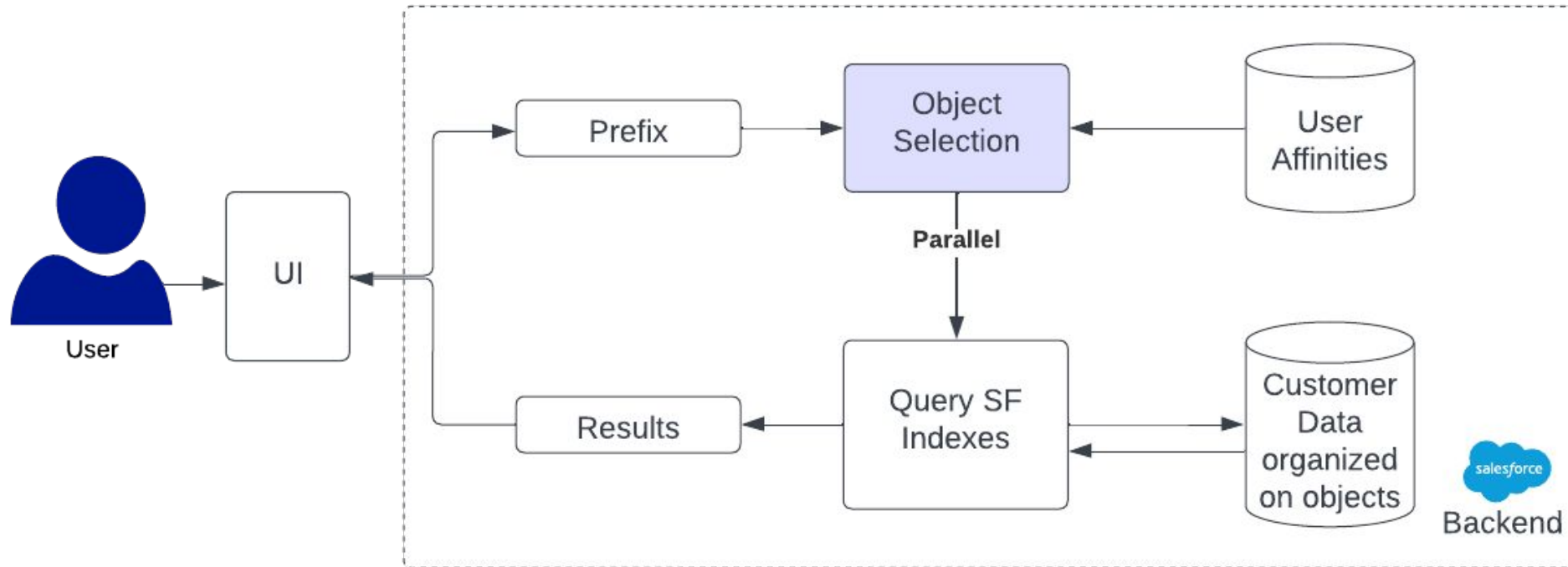
Salesforce customers are very diverse. Salesforce allows them to define hundreds of custom object types. For autocomplete scalability we need:

- To only query a subset of objects
- To scale to different domains
- To deal with new users/customers (cold-start)

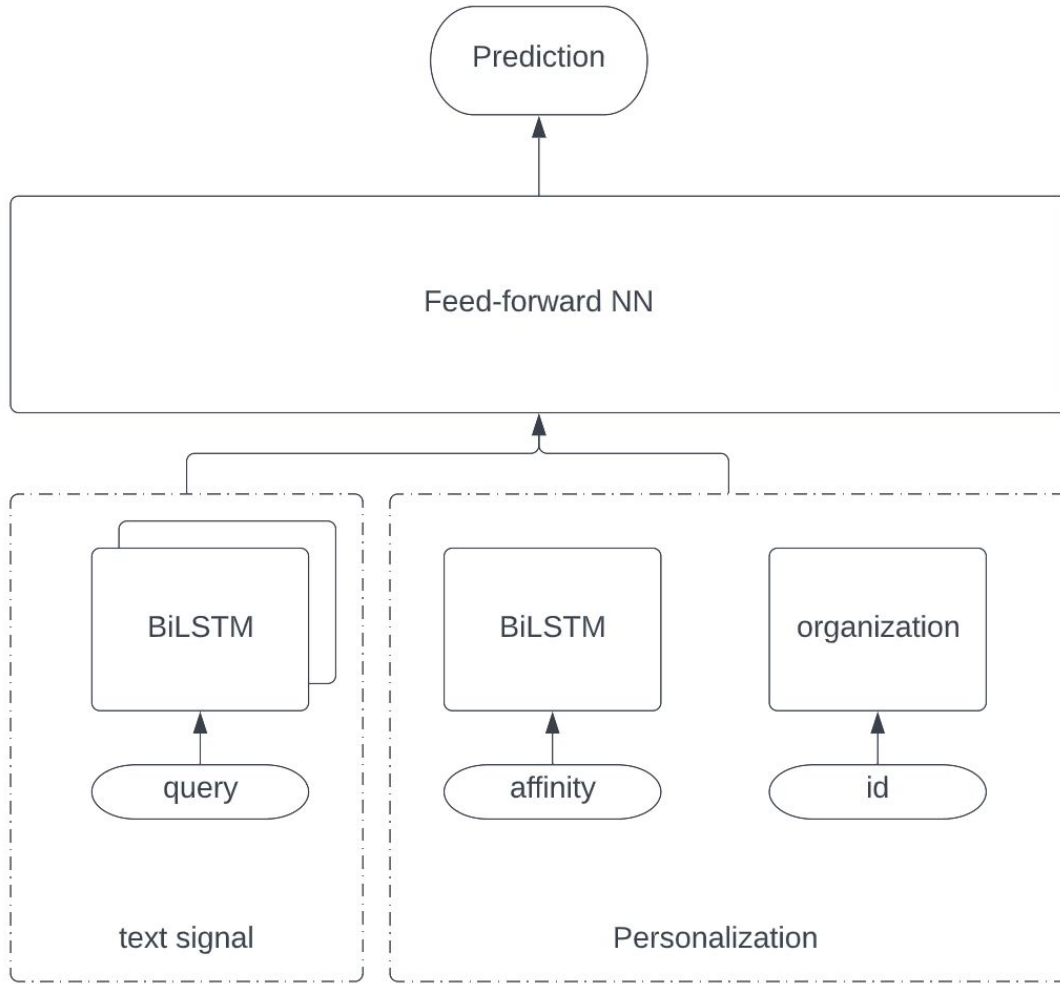# The legacy system builds user profiles based on clicks with spark jobs.

Advantages:
- Scalability
- Domain independent
- Low maintenance

Disadvantages:
- Query text independent
- Low discoverability of new objects (those out of user's affinity)

The deep learning solution uses these profiles and the query text.

Advantages:
- Query is an input
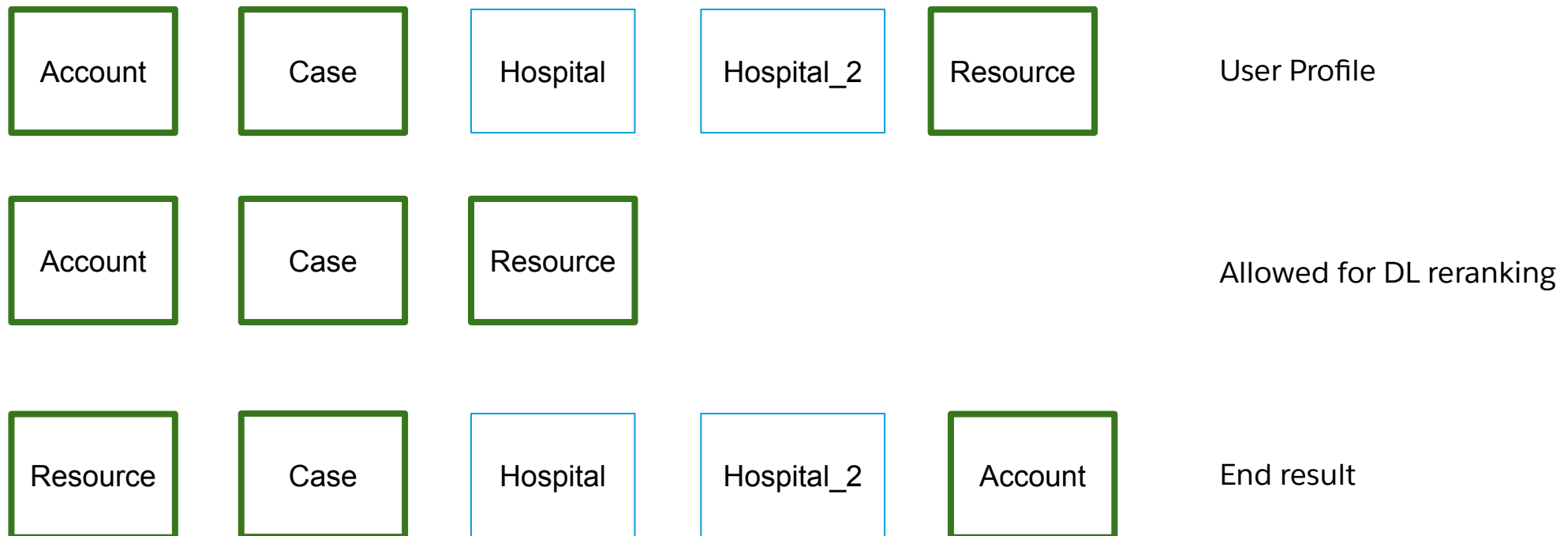- More complex and powerful decision function

Disadvantages:
- Explainability
- Distribution shift
- Maintenance + eng. cost to integrate
- Latency to get the model predictions

We use both in a hybrid solution that combines their advantages.

- Allow the DL model to re-rank only a subset of entities for each organization
- This subset is selected based on offline performance

| Account | Case | Hospital | Hospital_2 | Resource | User Profile |

| Account | Case | Resource | | | Allowed for DL reranking |

| Resource | Case | Hospital | Hospital_2 | Account | End result |

A/B metrics

- Implemented the model with ml4ir (https://github.com/salesforce/ml4ir)
- A/B tested vs the legacy solution
    - during March of 2021
    - thousands of users
    - millions of clicks
    - hundreds of organizations
- Latency
    - 15% improvement on the backend
    - Reducing cost-to-serve by 18% (CPU consumption)
- Relevance
    - Also improved as measured by the CTR of autocomplete suggestions

Concluding remarks

General

- ML models can improve several aspects of search systems
- They come with their set of challenges in terms of maintenance and integration

Data Collection

- Impacts the quality of the suggestions
- Domain adaptation: Autocomplete acts on query prefixes. Adapting training of the DNN model on this improved offline perf. significantly
- Domain adaptation: can be simple e.g., randomly sample prefixes or collect better training data

A/B test

- KPI and guardrail metrics need to be carefully designed in advance