Scaling up the Contacts Insights with Activity Graph

Praveen Innamuri, Zhidong Ke
Salesforce
Agenda

- Introduction
- Activity Insights Context
- Why using a Graph to model context
- Key problems solved and lessons learned
- Wrap up and QAs
Forward-Looking Statement

Statement under the Private Securities Litigation Reform Act of 1995

This presentation may contain forward-looking statements that involve risks, uncertainties, and assumptions. If any such uncertainties materialize or if any of the assumptions proves incorrect, the results of salesforce.com, inc. could differ materially from the results expressed or implied by the forward-looking statements we make. All statements other than statements of historical fact could be deemed forward-looking, including any projections of product or service availability, subscriber growth, earnings, revenues, or other financial items and any statements regarding strategies or plans of management for future operations, statements of belief, any statements concerning new, planned, or upgraded services or technology developments and customer contracts or use of our services.

The risks and uncertainties referred to above include – but are not limited to – risks associated with developing and delivering new functionality for our service, new products and services, our new business model, our past operating losses, possible fluctuations in our operating results and rate of growth, interruptions or delays in our Web hosting, breach of our security measures, the outcome of any litigation, risks associated with completed and any possible mergers and acquisitions, the immature market in which we operate, our relatively limited operating history, our ability to expand, retain, and motivate our employees and manage our growth, new releases of our service and successful customer deployment, our limited history reselling non-salesforce.com products, and utilization and selling to larger enterprise customers. Further information on potential factors that could affect the financial results of salesforce.com, inc. is included in our annual report on Form 10-K for the most recent fiscal year and in our quarterly report on Form 10-Q for the most recent fiscal quarter. These documents and others containing important disclosures are available on the SEC Filings section of the Investor Information section of our Web site.

Any unreleased services or features referenced in this or other presentations, press releases or public statements are not currently available and may not be delivered on time or at all. Customers who purchase our services should make the purchase decisions based upon features that are currently available. Salesforce.com, inc. assumes no obligation and does not intend to update these forward-looking statements.
Why I’m talking about Spotify:

- No, I’m not promoting to use Spotify.
- I should rather promote to use Salesforce products.
The Age of the Customer
Salesforce Apps + AI = Whole New Customer Experience

Sales Cloud
Predictive Lead Scoring
Opportunity Insights
Automated Activity Capture
AI Inbox

Commerce Cloud
Product Recommendations
Predictive Sort
Commerce Insights

App Cloud
Heroku + PredictionIO
Predictive Vision Services
Predictive Sentiment Services
Predictive Modeling Services

Analytics Cloud
Predictive Wave Apps
Smart Data Discovery
Automated Analytics & Storytelling

Service Cloud
Recommended Case Classification
Recommended Responses
Predictive Close Time

Marketing Cloud
Predictive Scoring
Predictive Audiences
Automated Send-time Optimization

Community Cloud
Recommended Experts, Articles & Topics
Automated Service Escalation
Newsfeed Insights

IoT Cloud
Predictive Device Scoring
Recommend Best Next Action
Automated IoT Rules Optimization

The Age of the Customer
Salesforce Apps + AI = Whole New Customer Experience
Augment CRM using AI and activity capture

- Automatic activity capture
- Emails, meetings, tasks, calls, etc.
- Extract Insights through classification
- Pricing discussed, Executive involved, Scheduling Requested, Product Mention, recommended connection etc.
- Context Activities, CRM, etc.
- AI Inbox
- Suggest Action(s)
- Timelines
- Other Salesforce Apps...

Extract Insights through classification.

- AI Inbox
- Suggest Action(s)
- Timelines
- Other Salesforce Apps...
Salesforce Apps - Closest Connections

Profile

Acme Health Inc
San Francisco

Show All Contact Sources

People Connected to Evan

Michael Machado
michael.machado@acmehealthinc.com
Because of emails and meetings they had... View

Richa Bansal
richa.bansal@acmehealthinc.com
Because of emails and meetings they had...

Sriram Iyer
sriram.iyer@acmehealthinc.com
Because of emails and meetings they had together.

Amy Miller
amy.miller@acmehealthinc.com
Because of emails and meetings they had...

View More

Call | SMS | Email
Agenda

• Introduction

• **Contact Insights Context**

• Why using a Graph to model context

• Key problems solved and lessons learned

• Wrap up and QAs
AI & Context

Data + Algorithms + Compute = Killer Apps

What does all those apps have in common? **User context**
Consumer vs Enterprise Context

User isn’t the product but the customer
  • Retention, privacy, GDPR, security, auditing, etc

Context has to be scoped
  • Cannot be used globally: organization, team, user levels

Very rich
  • Goes way beyond user context: organizations/groups/teams, products and services, companies, different types of activities across many different products, etc

Very dynamic
  • Fast coming data with lots of interaction points
Context enables us to deliver deeper insights.

Go beyond using a single email to make classification and action recommendation

This sender looks familiar, how well should I know him / her?
  • Are we strongly connected? Is he or she important to my accounts or opportunities? etc

Is this email discussing products or services that my company sell?

Is this email discussing competitors?

Who, in my org, can help me sell to an individual or company?
  • Supply relevant background information on a particular individual or company
  • Identify who is the key decision maker
  • Give me historical information for that individual or company
  • Make an introduction for me
Agenda

• Introduction
• Activity Insights Context
• **Why using a Graph to model context**
• Key problems solved and lessons learned
• Wrap up and QAs
A graph is an efficient means for encoding relationships.

An org can have thousands of contacts
- These contacts exist within the org itself (e.g., sales rep, account exec)
- Perhaps more importantly, contacts extend beyond the org (e.g., buyers)

That same org can have millions of events per week
- Events (e.g., meetings, emails, phone calls) connect contacts and indicate a relationship
- The number and nature of events between contacts can indicate strength of connection / relationship
Coupled with AI models, our graph delivers Contextual Services.

Graph

- U
- A
- B
- T
- D
- C

Who is a particular email from and why should I care? Role, latest communication, meeting history, mutual friends, contact info, etc.

Context

Insights

- Identify hot leads
- Best time to email
- Recommend connections
- Updated contact info notification
- Suggest recipients, or rooms, for meetings
- Identify contact’s role: economic buyer, evaluator, influencer, etc.
- Relationship with contact: e.g., strength of connection, communication topics

Models

- Pricing discussed
- Scheduling requested
- Exec involved
- etc.
High Level graph generation architecture

Activity Stream

kafka

Apache Spark

Activity store

Index Store

API Layer

Clients

File Store

Persist/load

Index

Bootstrap

Delivery of Graph Services

Computing Graph Services
Activity events to create / update the raw graph.

Graph Update
(Using Spark GraphX)

SAS Δ-batch → RDD[Activity]

File Store → RDD(VertexId, Contact)

Old Raw Graph

Participants

Consolidate as (VertexId, Contact)

Consolidate as (Edge[Events])

Merge edge and vertex RDDs

Updated Raw Graph

RDD((EdgeKey, Events))

File Store
Architecture Diagram - Onboarding for all Orgs

- Records Store
- Activity store
- Driver
- Executor
- Executor
- Executor
- Executor

- Load Activity Data
- Graph Generation
- Graph checkpoint
- Load into File Store

- Load Graph data
- Compute Insights
- Persist/Index those insights
Agenda

• Introduction
• AI & context
• Why using a Graph to model context
• Problems & Lessons Learned
• Wrap up and QAs
Memory Issues

java.lang.OutOfMemoryError: Java heap space
  at java.util.Arrays.copyOf(Arrays.java:3236) ~[na:1.8.0_121]
  at java.io.ByteArrayOutputStream.grow(ByteArrayOutputStream.java:118)
  at java.io.ByteArrayOutputStream.ensureCapacity(ByteArrayOutputStream.java:93)
  at java.io.ByteArrayOutputStream.write(ByteArrayOutputStream.java:153)
Bucketing Strategy

Org Graph Jobs

Group By Size

- Small Bucket
- Medium Bucket
- Large Bucket
- Extra Large Bucket
- ...

In-Memory Computing Summit North America 2018
Tuning

Find Right Memory, #Executors and #Partitions Per Bucket

Created by blog.3back.com
Spark Job Got Stuck Before Reading Data

val df = spark.read.avro("input/* .avro")

Too many small files =>
Solution

Bypass the metadata fetch

```
val df = spark.read.avro("input/\*.avro")
val input = sc.parallelize(List("input/"), 10).map(_.readData)
```
Compaction Framework

Compact small files

One spark job for a batch of requests
Scaling

How to scale up from 0 -> Thousands of orgs?

Created by Freepik
Scale Up #Clusters

Hash partition org within each bucket
Spin up multiple spark clusters
“Hotspot” issue
Solution

Create a request queue for each bucket
Some Useful Tips

- Bucketing Strategy for variant data input
- Partition the orgs into small bins within each bucket
- Try scale up with multiple spark clusters
- Say no to tiny files and compact them to large chunk
- Use a simple queue with pulling module can balance the load
Re-compute full graph or Incremental updates
Incremental update

- Save the intermediate graph data and checkpoint
- Incremental updating the contacts
Failure happens

<table>
<thead>
<tr>
<th>sparkJob</th>
<th>2018-08-20 10:36 55s</th>
<th>2018-08-22 19:00 40s</th>
<th>2d 8h 23m</th>
<th>Success</th>
<th>Details</th>
</tr>
</thead>
<tbody>
<tr>
<td>sparkJob</td>
<td>2018-08-20 10:36 55s</td>
<td>2018-08-22 17:52 59s</td>
<td>2d 7h 16m</td>
<td>Success</td>
<td>Details</td>
</tr>
<tr>
<td>sparkJob</td>
<td>2018-08-20 10:36 55s</td>
<td>2018-08-23 08:06 36s</td>
<td>2d 2h 29m</td>
<td>Success</td>
<td>Details</td>
</tr>
<tr>
<td>sparkJob</td>
<td>2018-08-20 10:36 55s</td>
<td>2018-08-22 21:07 41s</td>
<td>2d 10h 30m</td>
<td>Success</td>
<td>Details</td>
</tr>
<tr>
<td>sparkJob</td>
<td>2018-08-20 10:36 55s</td>
<td>2018-08-23 02:28 17s</td>
<td>2d 15h 51m</td>
<td>Success</td>
<td>Details</td>
</tr>
<tr>
<td>sparkJob</td>
<td>2018-08-20 10:36 55s</td>
<td>2018-08-22 15:36 59s</td>
<td>2d 5h 0m</td>
<td>Success</td>
<td>Details</td>
</tr>
<tr>
<td>sparkJob</td>
<td>2018-08-20 10:42 32s</td>
<td>2018-08-20 10:43 04s</td>
<td>32 sec</td>
<td>Failed</td>
<td>Details</td>
</tr>
<tr>
<td>sparkJob</td>
<td>2018-08-21 05:58 26s</td>
<td>2018-08-21 05:58 26s</td>
<td>0 sec</td>
<td>Cancelled</td>
<td>Details</td>
</tr>
</tbody>
</table>

one failed job with many succeed jobs

<table>
<thead>
<tr>
<th>264</th>
<th>2018/10/02 17:27:36</th>
<th>85 ms</th>
<th>2/2 (5 skipped)</th>
<th>6/6 (20 skipped)</th>
</tr>
</thead>
<tbody>
<tr>
<td>263</td>
<td>2018/10/02 17:27:33</td>
<td>2 s</td>
<td>3/3 (3 skipped)</td>
<td>12/12 (12 skipped)</td>
</tr>
<tr>
<td>262</td>
<td>2018/10/02 17:27:33</td>
<td>95 ms</td>
<td>2/2 (2 skipped)</td>
<td>8/8 (8 skipped)</td>
</tr>
<tr>
<td>251</td>
<td>2018/10/02 17:27:33</td>
<td>24 ms</td>
<td>1/1 (2 skipped)</td>
<td>4/4 (8 skipped)</td>
</tr>
<tr>
<td>250</td>
<td>2018/10/02 17:27:33</td>
<td>24 ms</td>
<td>1/1 (1 skipped)</td>
<td>3/3 (3 skipped)</td>
</tr>
<tr>
<td>258</td>
<td>2018/10/02 17:27:33</td>
<td>25 ms</td>
<td>1/1 (5 skipped)</td>
<td>4/4 (3 skipped)</td>
</tr>
<tr>
<td>257</td>
<td>2018/10/02 17:27:33</td>
<td>18 ms</td>
<td>1/1</td>
<td>25/25 (8 skipped)</td>
</tr>
<tr>
<td>256</td>
<td>2018/10/02 17:27:33</td>
<td>14 ms</td>
<td>1/1</td>
<td>3/3</td>
</tr>
<tr>
<td>255</td>
<td>2018/10/02 17:27:33</td>
<td>14 ms</td>
<td>1/1</td>
<td>1/1</td>
</tr>
</tbody>
</table>

a lot of stages and jobs for graph generations
Failure Recover ?

Check State ?
Metadata Store

Metadata Store and Retry pipeline

Activity Data Lake → Onboarding

Graph Data Lake → Incremental activity from last checkpoint

Metadata Database

Checkpoint ↔ Metadata Database ↔ retry [Apache Spark]
Indexing failures

Corruption Problems
Index updates

Day-0 Job

Incremental Job

Group by Org

Group by Org

Index-10/02/2018-10:00

Correction

Index-10/04/2018-10:00

Validate

API Server

Succeed

Failed

Succeed
Some Lessons

- Use Incremental updates
- Create a metadata table for checkpoint and state store
- Create Indexes for each iteration of contact insights
Agenda

• Introduction
• Activity Insights Context
• Why using a Graph to model context
• Key problems solved and lessons learned
• **Wrap up and QAs**
Future Work

- Explore the graph database
- Explore the in-memory database Apache Ignite
THANK YOU

salesforce.com/careers