Dropbox Analytics
Steve Bartel
Talk Structure

1. Introduction to Dropbox
2. Analytics at Dropbox
3. Data Set
4. Future
5. Q&A
Talk Structure

1. Introduction to Dropbox
2. Analytics at Dropbox
3. Data Set
4. Future
5. Q&A
What is Dropbox?
Talk Structure

1. Introduction to Dropbox
2. Analytics at Dropbox
3. Data Set
4. Future
5. Q&A
Dropbox Analytics Goals

• Make Dropbox a data-driven company
  – Inform decisions
    • Justify new features
    • Inform business deals
    • Respond to abuse
  – Monitor core metrics over time
    • Health of different platforms & features
    • Track growth, engagement, financials
Dropbox Analytics Functions
Dropbox Analytics Functions

• Analysis
  – Pre and Post shmodel analytics
  – Android usage to justify HTC deal
  – Camera upload abuse
Dropbox Analytics Functions

• Analysis
  – Pre and Post shmodel analytics
  – Android usage to justify HTC deal
  – Camera upload abuse

• Tools / dashboards
Dropbox Analytics Functions

• Analysis
  – Pre and Post shmodel analytics
  – Android usage to justify HTC deal
  – Camera upload abuse
• Tools / dashboards
• Infrastructure
Talk Structure

1. Introduction to Dropbox
2. Analytics at Dropbox
3. Data Set
4. Future
5. Q&A
Data Set - Contents
Data Set - Contents

• User metadata
Data Set - Contents

• User metadata
• File metadata
What takes up the most space after 3 weeks of upgrading?

• Music
• Docs
• Images
• Videos
Data Set - Contents

• User metadata
• File metadata
• File interaction
Data Set - Contents

• User metadata
• File metadata
• File interaction
• Sharing / referral graph
Data Set - Scale

- Petabytes of file data
- 500 million files saved to dropbox each day
- Several billion new analytics records every day
Talk Structure

1. Introduction to Dropbox
2. Analytics at Dropbox
3. Data Set
4. Future
5. Q&A
Future

• Photos
• Classification
• Security/Fraud
We’re hiring!

• Dropbox Engineering Team – 55 people
• Engineering Analytics Team – 5 people
• Data scientists – 0 people
• Email me at sbartel@dropbox.com
• dropbox.com/jobs
Talk Structure

1. Introduction to Dropbox
2. Analytics at Dropbox
3. Data Set
4. Future
5. Q&A
Discussion / Questions

• About
  – Dropbox
  – Analysis @ Dropbox
  – Analytics Architecture / Stack
  – Roadmap
Metaservers

Prod MySQL databases

Squoop

cLogger

distcp

Hadoop

Rollups via Hive queries

Hive query

Dashboard Cache

Dashboards

analytics.dropbox.com/hive
Metaservers → cLogger → distcp → Hadoop
Prod MySQL databases → Sqoop → Hadoop
Hive query → Hive queries → Rollups via Hive queries
Dashboard Cache → Dashboards
analytics.dropbox.com/hive