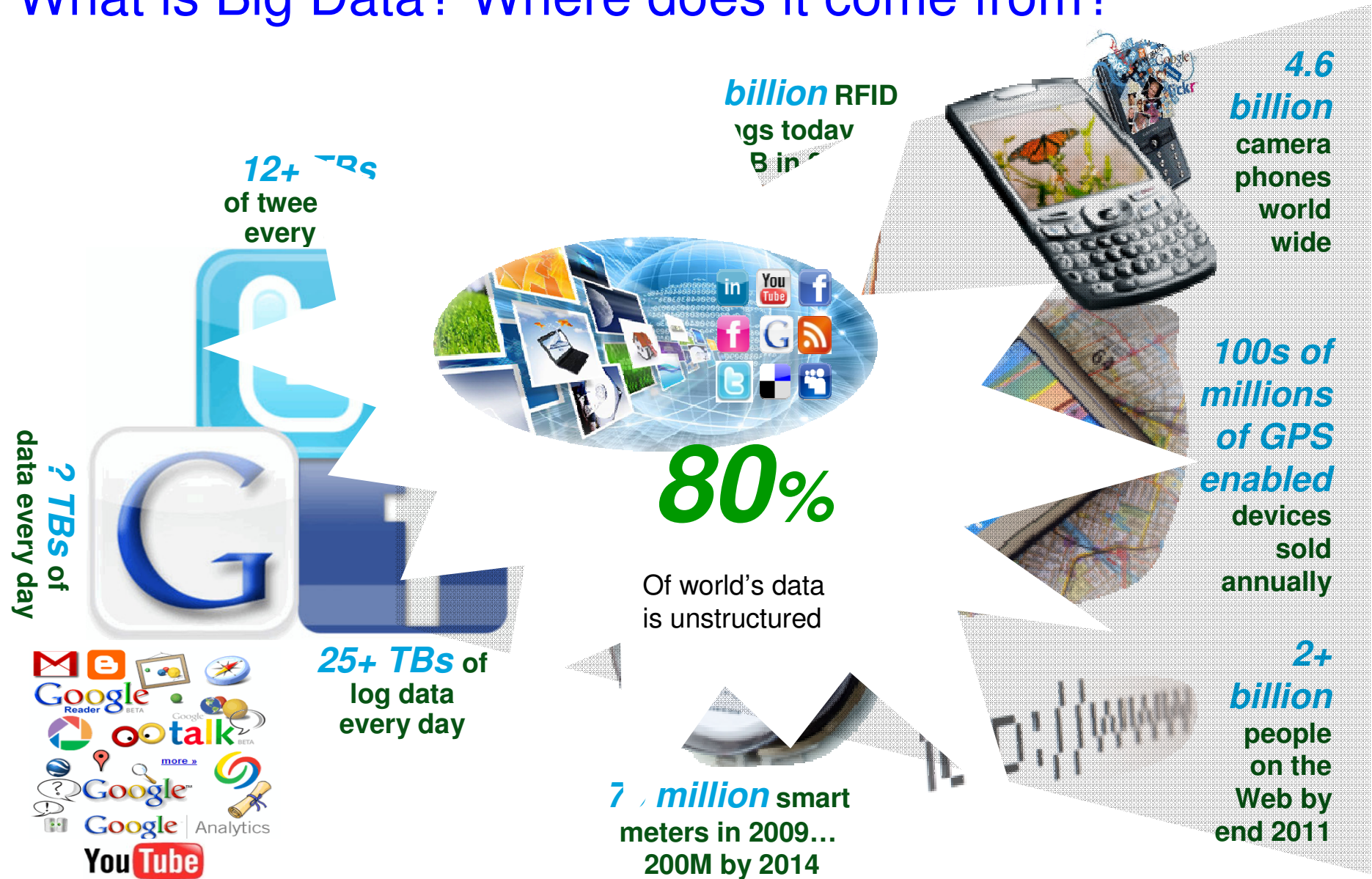# Exploiting Data at Rest and Data in Motion with a Big Data Platform

Sarah Brader, sarah_brader@uk.ibm.com

What is Big Data? Where does it come from?
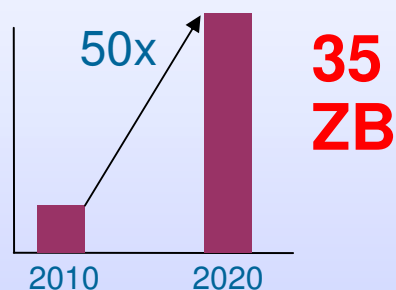
billion RFID
~~gs today~~
~~B in~~

4.6 billion camera phones world wide

12+ TBs
of twee
every

? TBs of data every day

80%

Of world's data is unstructured

100s of millions of GPS enabled devices sold annually

25+ TBs of log data every day

7 million smart meters in 2009…
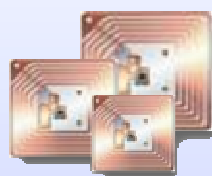200M by 2014

2+ billion people on the Web by end 2011

Characteristics of Big Data

Extracting insight from an immense volume, variety and velocity of data, in context, beyond what was previously possible.

Cost efficiently processing the growing **Volume**
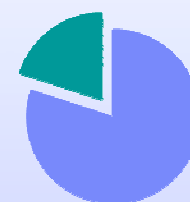
50x

35 ZB

2010　2020

Responding to the increasing **Velocity**

30 Billion RFID sensors and counting
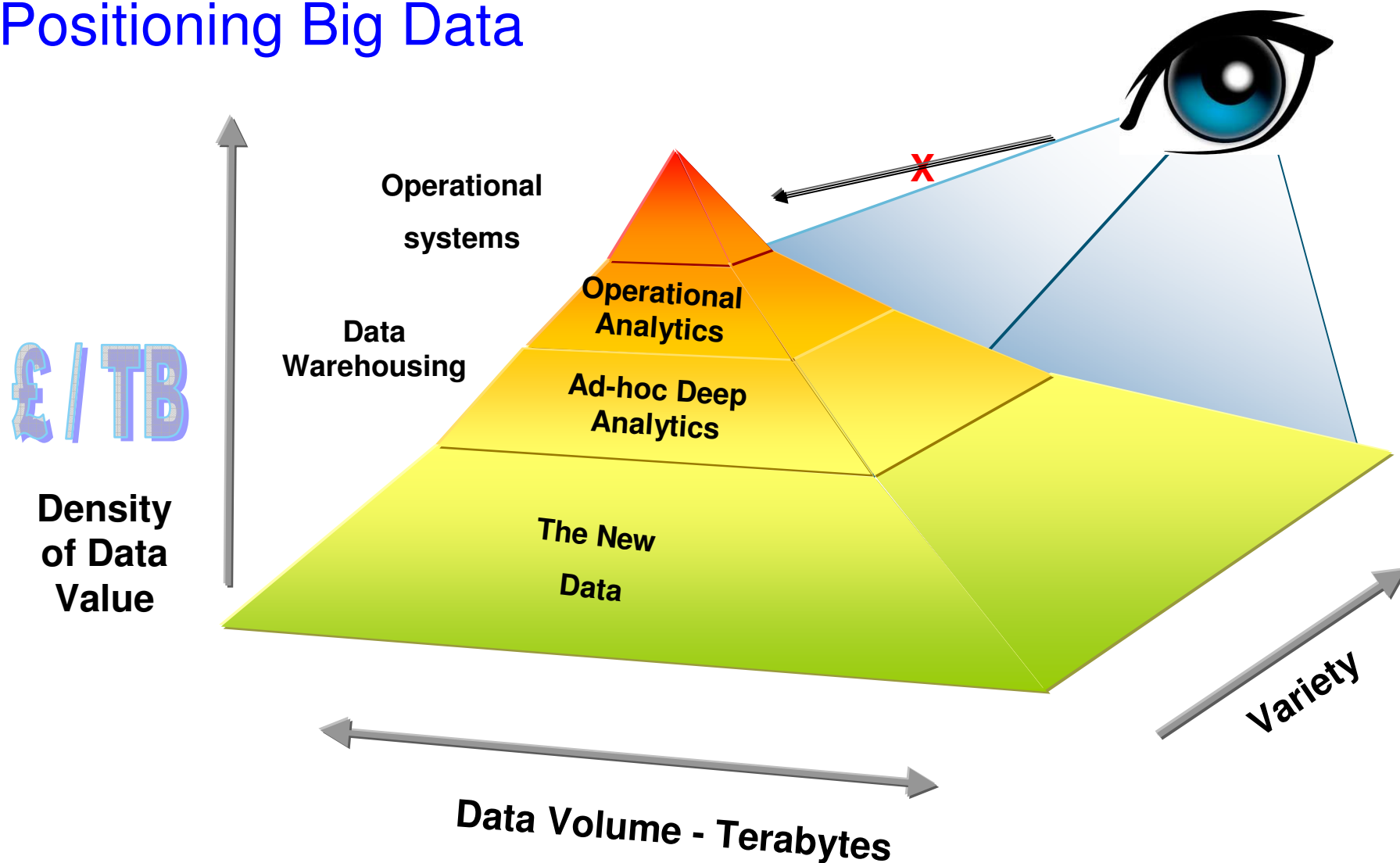
Collectively Analyzing the broadening **Variety**
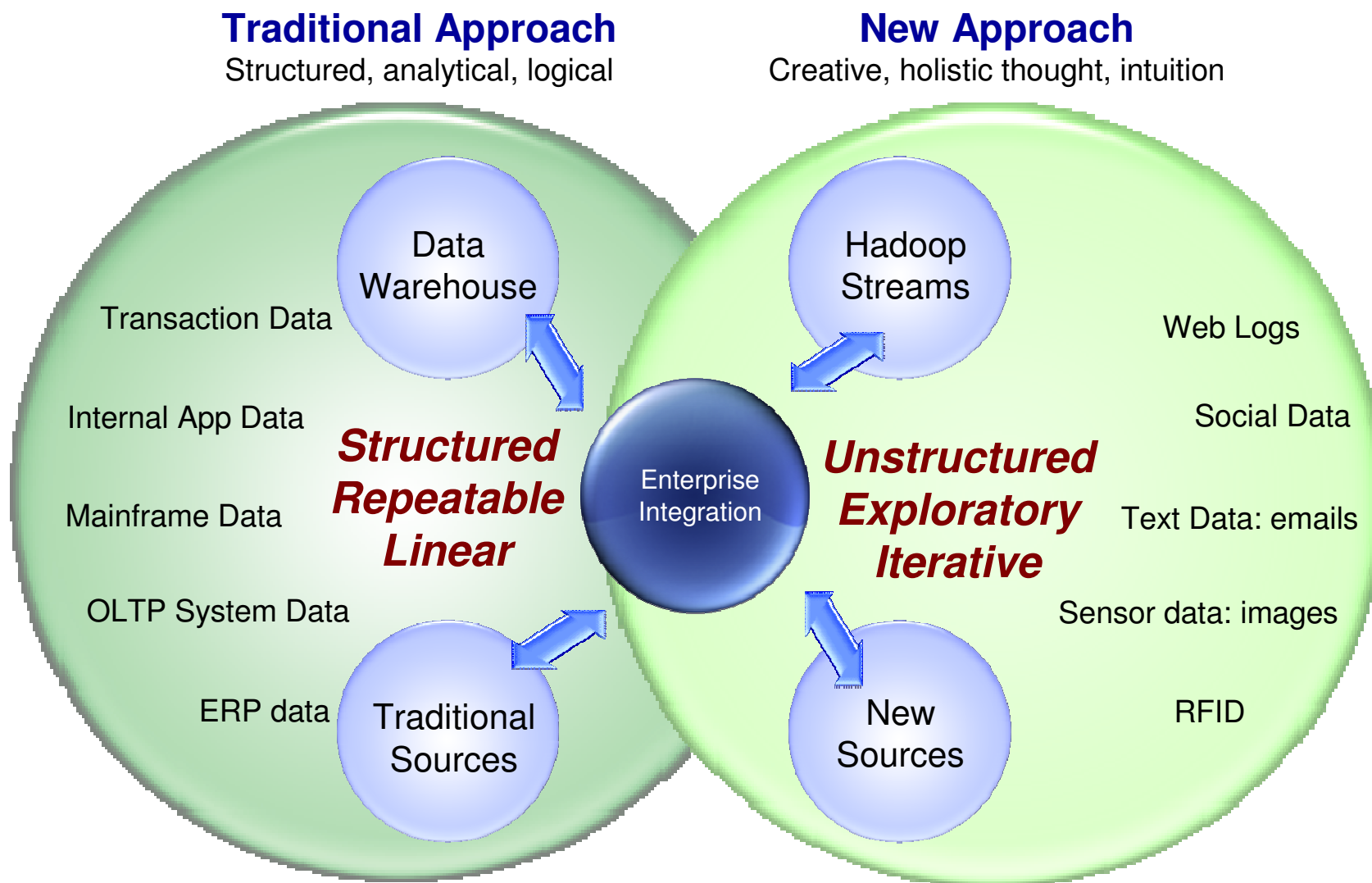
80% of the worlds data is unstructured

Establishing the **Veracity** of big data sources

1 in 3 business leaders don't trust the information they use to make decisions
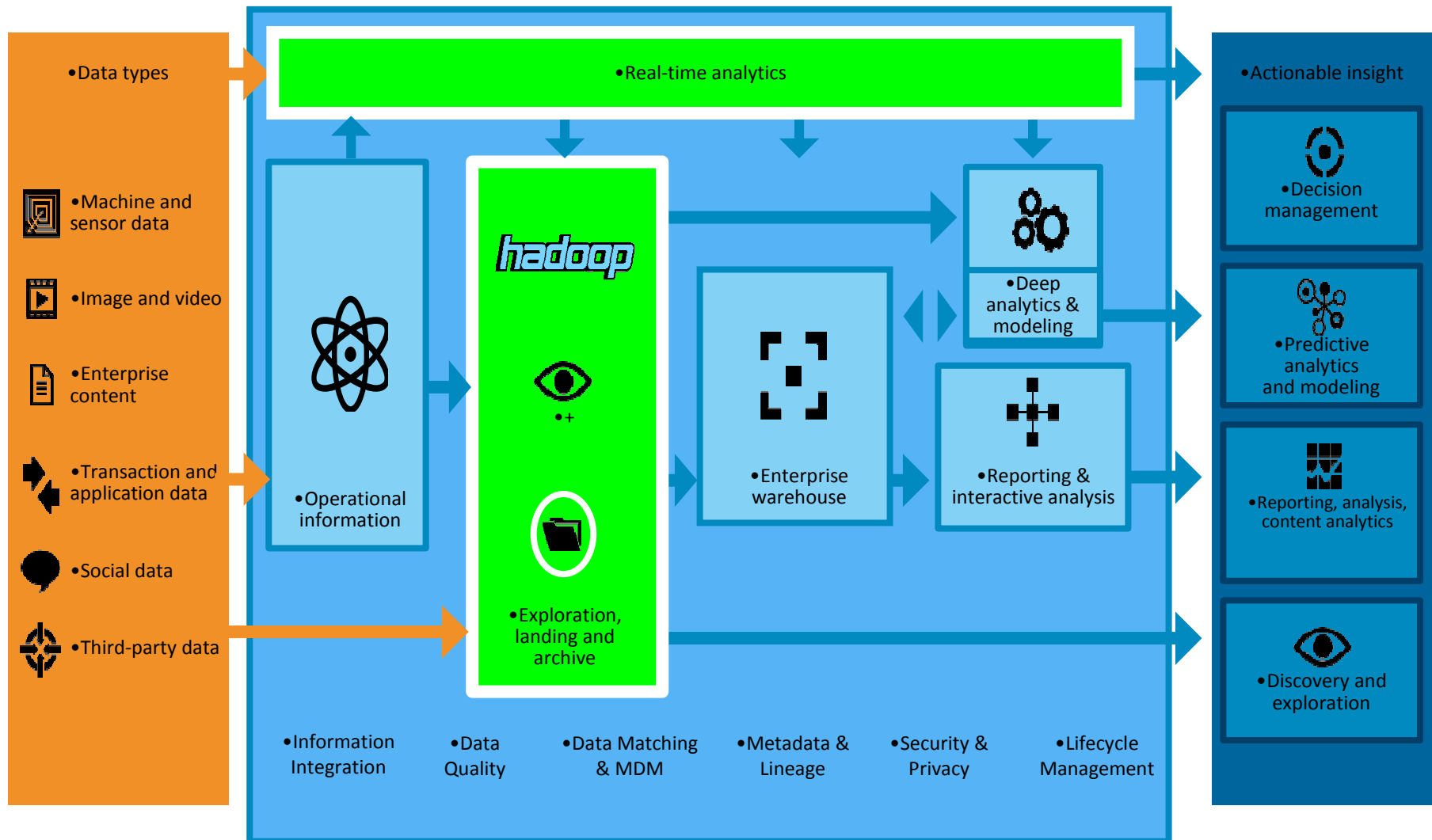
Positioning Big Data

£ / TB

Density of Data Value

Operational systems

Data Warehousing

Operational Analytics

Ad-hoc Deep Analytics

The New Data

X

Variety

Data Volume - Terabytes

A New Approach for Big Data

Traditional Approach
Structured, analytical, logical

New Approach
Creative, holistic thought, intuition

Data Warehouse

Hadoop Streams

Transaction Data

Web Logs

Internal App Data

Structured Repeatable Linear

Social Data

Enterprise Integration

Unstructured Exploratory Iterative

Mainframe Data

Text Data: emails

OLTP System Data

Sensor data: images

ERP data

Traditional Sources

New Sources

RFID

IBM's Next Generation Big Data Architecture

Data types

- Machine and sensor data
- Image and video
- Enterprise content
- Transaction and application data
- Social data
- Third-party data

- Real-time analytics

- Operational information

hadoop

- Exploration, landing and archive

- Enterprise warehouse

- Deep analytics & modeling

- Reporting & interactive analysis

- Information Integration
- Data Quality
- Data Matching & MDM
- Metadata & Lineage
- Security & Privacy
- Lifecycle Management

- Actionable insight
  - Decision management
  - Predictive analytics and modeling
  - Reporting, analysis, content analytics
  - Discovery and exploration
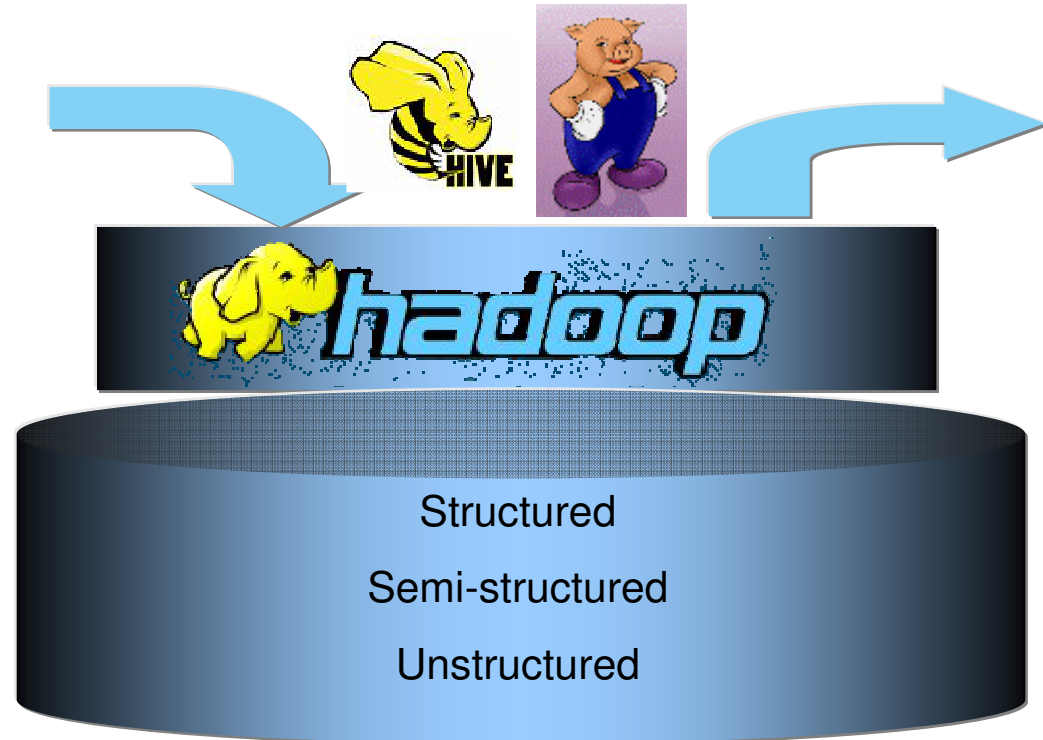
Hadoop Technology

- Volume
  - Scales to many petabytes
  - Scales to thousands of nodes

- Variety
  - Structured
  - Semi-structured
  - Unstructured

- Experimental analytics
  - Apply structure on read
  - Rapid exploitation of new data

Structured

Semi-structured

Unstructured

Hadoop ++   InfoSphere BigInsights

- Based on Apache Open Source
  - Integrates with OEM tech
  - Applications developed elsewhere run without change

- Ease of use
  - Skills in SQL, R etc are instantly usable
  - Complexities of MR are hidden

- Enterprise capabilities
  - Security
  - Disaster Recovery
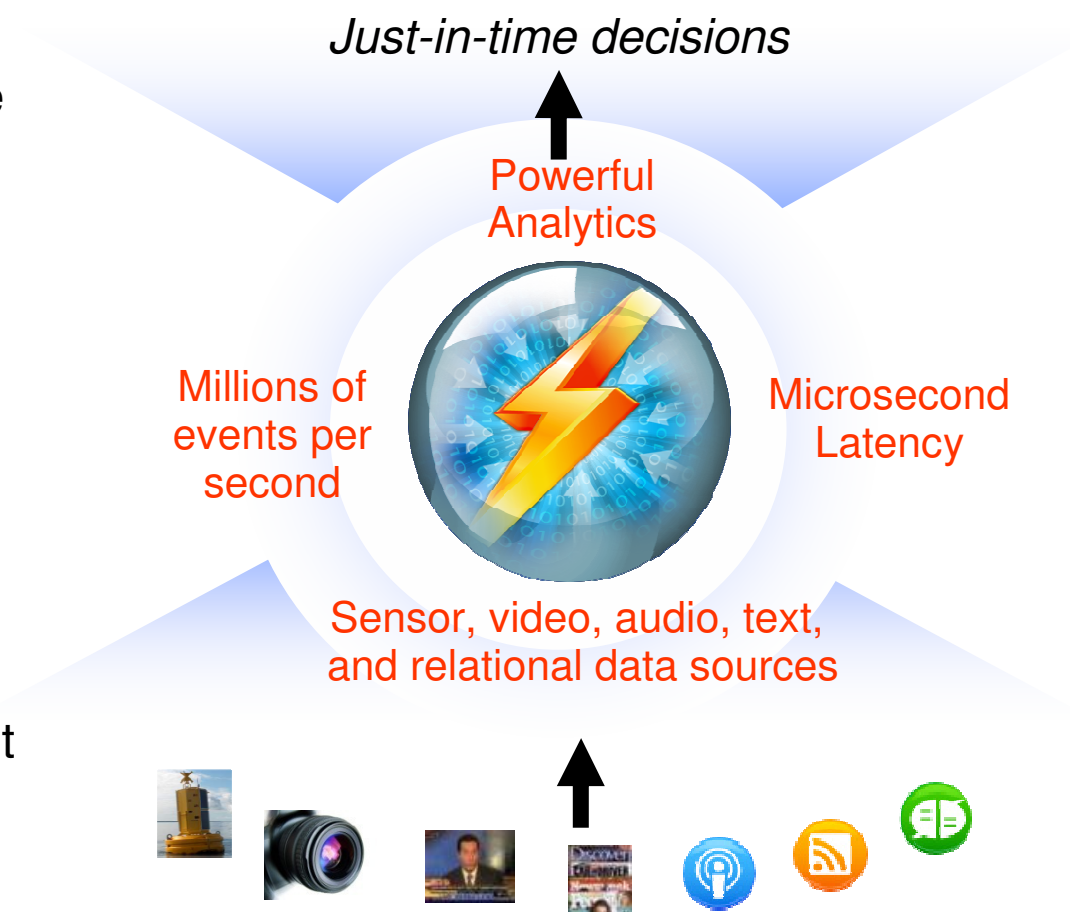 - Supports workload isolation

IBM InfoSphere Streams

A platform for real-time analytics on BIG data

- **Volume**
  - Gigabytes per second or more
  - Terabyte per day or more

- **Variety**
  - All kinds of data
  - All kinds of analytics

- **Velocity**
  - Insights in microseconds

- **Agility**
  - Dynamically responsive
  - Rapid application development

Just-in-time decisions

Powerful Analytics

Millions of events per second

Microsecond Latency

Sensor, video, audio, text, and relational data sources

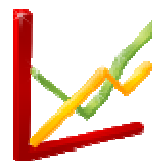# What kinds of analysis is Streams used for?

Telephony
- CDR processing
- Social analysis
- Churn prediction
- Geomapping

Stock market
- Impact of weather on securities prices
- Analyze market data at ultra-low latencies

Transportation
- Intelligent traffic management

Law Enforcement, Defense & Cyber Security
- Real-time multimodal surveillance
- Situational awareness
- Cyber security detection

Fraud prevention
- Detecting multi-party fraud
- Real time fraud prevention

Smart Grid & Energy
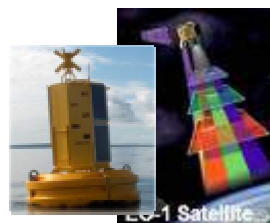- Transactive control
- Phasor Monitoring Unit

e-Science
- Space weather prediction
- Detection of transient events
- Synchrotron atomic research

Health & Life Sciences
- Neonatal ICU monitoring
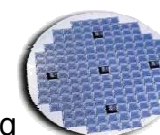- Epidemic early warning system
- Remote healthcare monitoring

Natural Systems
- Wildfire management
- Water management

Other
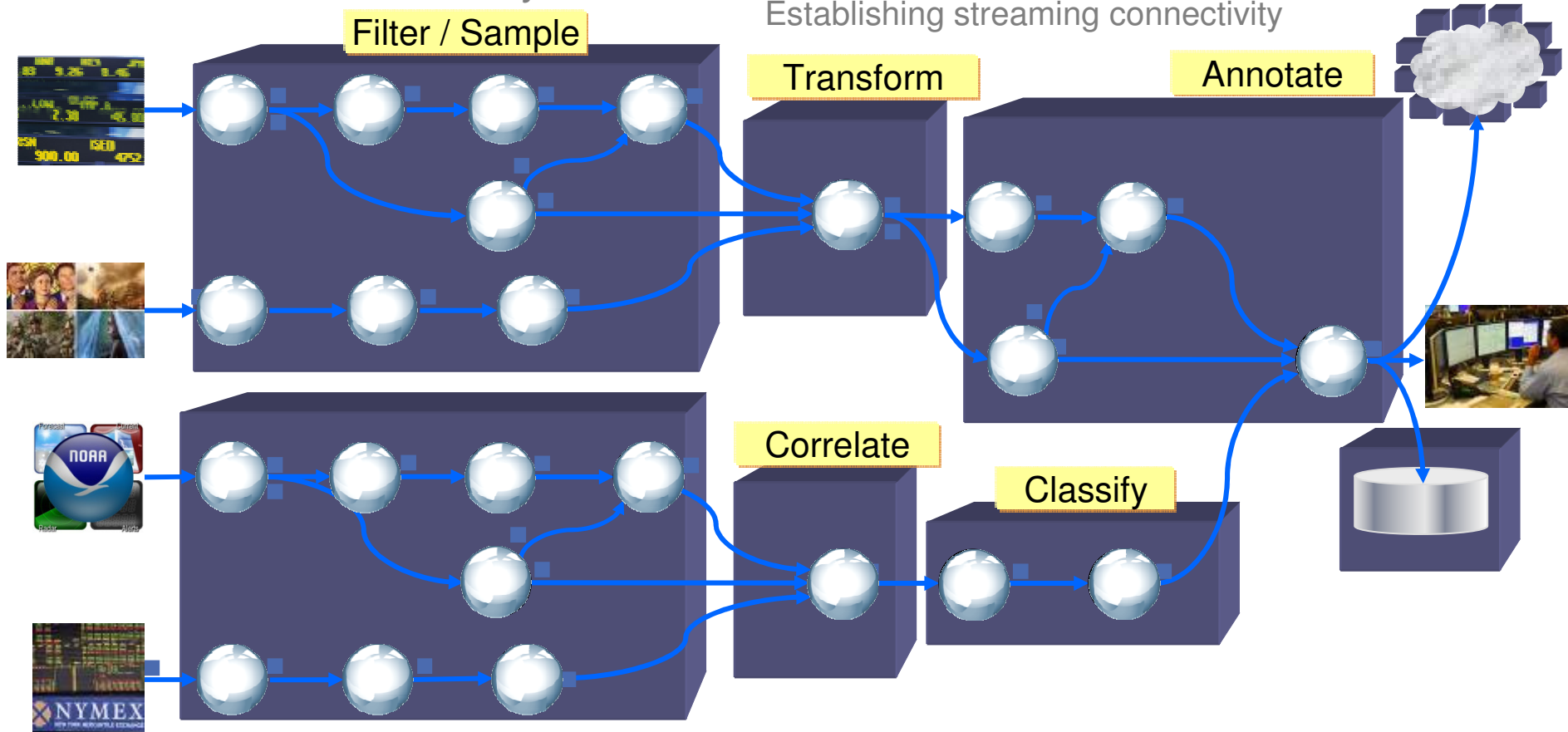- Manufacturing
- Text Analysis

How Streams Works

→ Continuous ingestion
→ Continuous analysis

How Streams Works

→ Continuous ingestion
    → Continuous analysis

Infrastructure provides services for
    Scheduling analytics across hardware hosts,
    Establishing streaming connectivity

Filter / Sample

Transform

Annotate

Correlate

Classify

Achieve scale:
    By partitioning applications into software components
    By distributing across stream-connected hardware hosts

Where appropriate:
    Elements can be *fused* together
    for lower communication latency

Further Information

- Resources
    - Visit IBMBigDataHub for conversations
        - www.ibmbigdatahub.com
    - Visit BigDataUniversity for technical education
        - www.bigdatauniversity.com
    - InfoSphere BigInsights 2.1.2 Knowledge Center
        - http://www-01.ibm.com/support/knowledgecenter/SSPT3X_2.1.2/com.ibm.swg.im.infosphere.biginsights.welcome.doc/doc/welcome.html
    - InfoSphere Streams Knowledge Center
        - http://www-01.ibm.com/support/knowledgecenter/SSCRJU/SSCRJU_welcome.html?lang=en
    - Watson Explorer Knowledge Center
        - http://www-01.ibm.com/support/knowledgecenter/SS8NLW/SS8NLW_welcome.html?lang=en

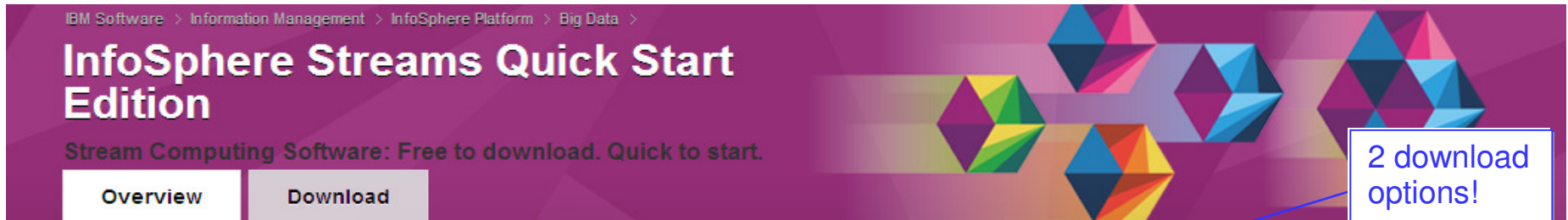# Try it now! InfoSphere for BigInsights Quick Start

- Free, no limit, non-production version of BigInsights
- Features Big SQL, BigSheets, Text Analytics, Big R, management console, development tools
- Tutorials and education available
- http://www.ibm.com/developerworks/downloads/im/biginsightsquick/

How Can You Get Quick Start?

Ibm.co/streamsqs

http://www.ibm.com/software/data/infosphere/streams/quick-start/

IBM Software > Information Management > InfoSphere Platform > Big Data >

InfoSphere Streams Quick Start Edition

Stream Computing Software: Free to download. Quick to start.

| Overview | Download |

2 download options!

Real Time Analytic Processing at your fingertips

Stream processing technology gives you the ability to analyze massive data volumes quickly, often in real time, and turn data into actionable insight.

InfoSphere Streams is an advanced computing platform that can quickly ingest, analyze and correlate information as it arrives from thousands of real-time sources. Because it can handle high throughput rates, InfoSphere Streams can analyze millions of events per second, enabling sub-millisecond response times and instant decision-making.

Now you can get your hands on this technology with InfoSphere Streams Quick Start Edition, a **free**, downloadable, non-production version.

With InfoSphere Streams Quick Start Edition, there is **no data capacity** and **no time limitation**, so you can experiment with large data sets and work with different use cases, on your own timeframe.

NOTE: InfoSphere Streams Quick Start Edition does not come with a support option. To explore your support options, visit the InfoSphere Streams product page.

Download now!

↓ **InfoSphere Streams Quick Start Edition Native Software Installation**

↓ **InfoSphere Streams Quick Start Edition VMware Image**

There are 2 download options available. Choose the native software option to download Quick Start onto your machine. Choose the VMware image for a virtual machine option to get started with Quick Start even faster.

Blogs

→ IBM developerWorks Big Data

→ IBM developerWorks Streams

→ Streams Blog on developerWorks

→ Stream Processing on Smarter Computing

→ IBM Big Data hub