Natural Sparksmanship

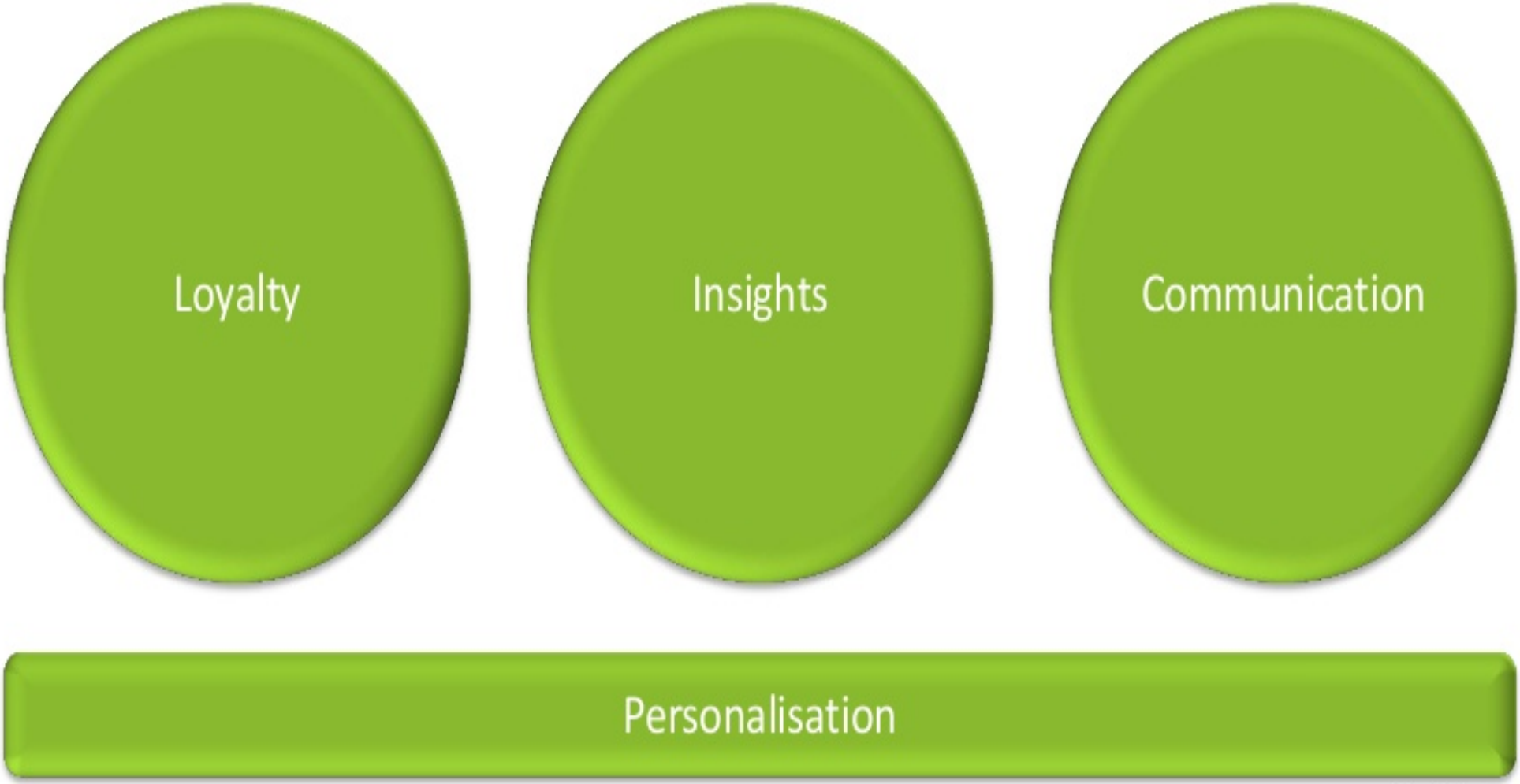
Art of Making an Analytics Enterprise cross the chasm

Rajesh Krishnan

AIMIA Inc.

SPARK SUMMIT 2016
DATA SCIENCE AND ENGINEERING AT SCALE

Our Business



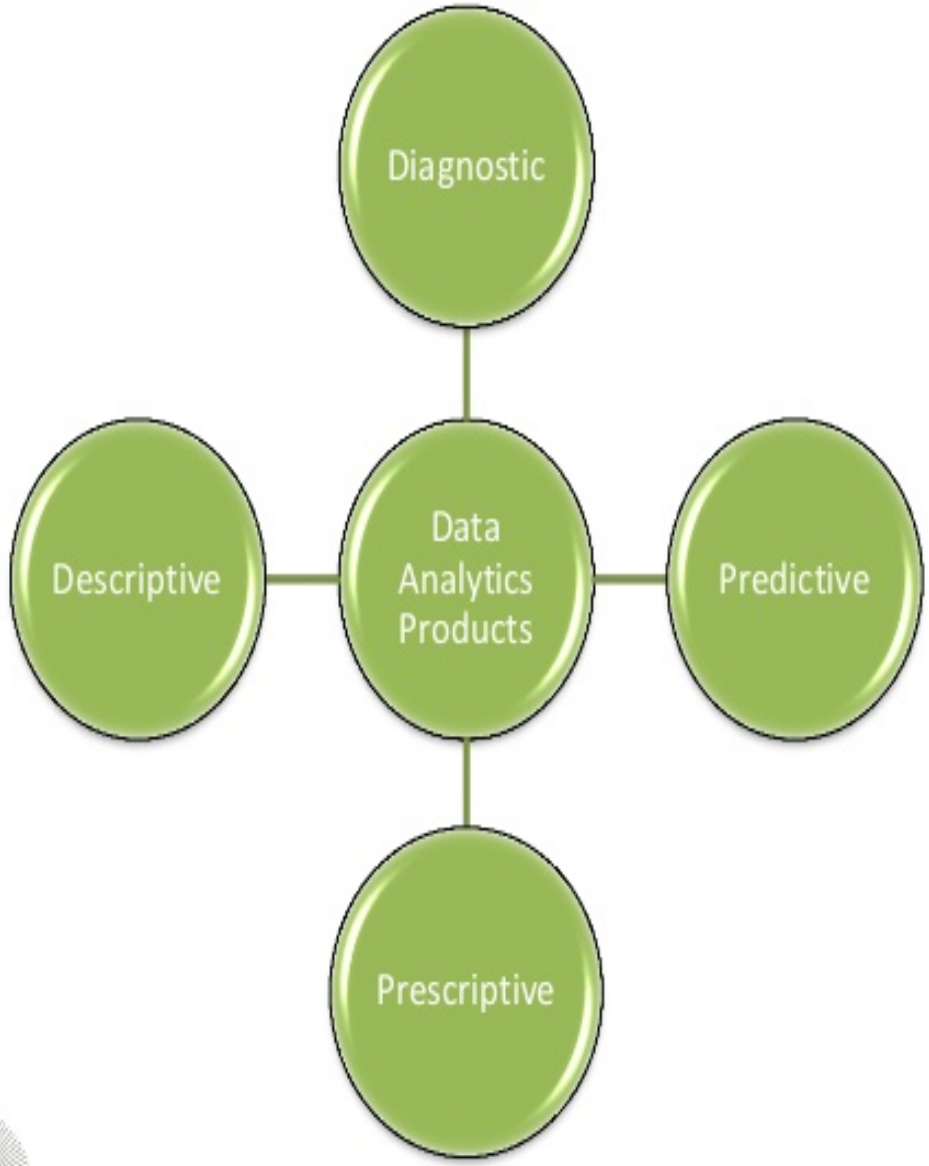
Our Business - notes

- AIMIA is data driven marketing & loyalty analytics business, regarded as a leader by Forrester in their loyalty management wave.
  - We manage UK's largest coalition **loyalty** program and also manage proprietary loyalty for customers across verticals and across geography
  - We provide **insights** through advanced analytics to retailers & CPGs across the globe.
  - We manage marketing & smart customer **communications** through campaigns for our clients utilising the analytics to drive loyalty.

- Personalisation underpins the three focus areas.

- Huge influx of data today means there is opportunity to know customer better. Personalisation can be done like never before.

Our Analytics



Our Analytics - notes

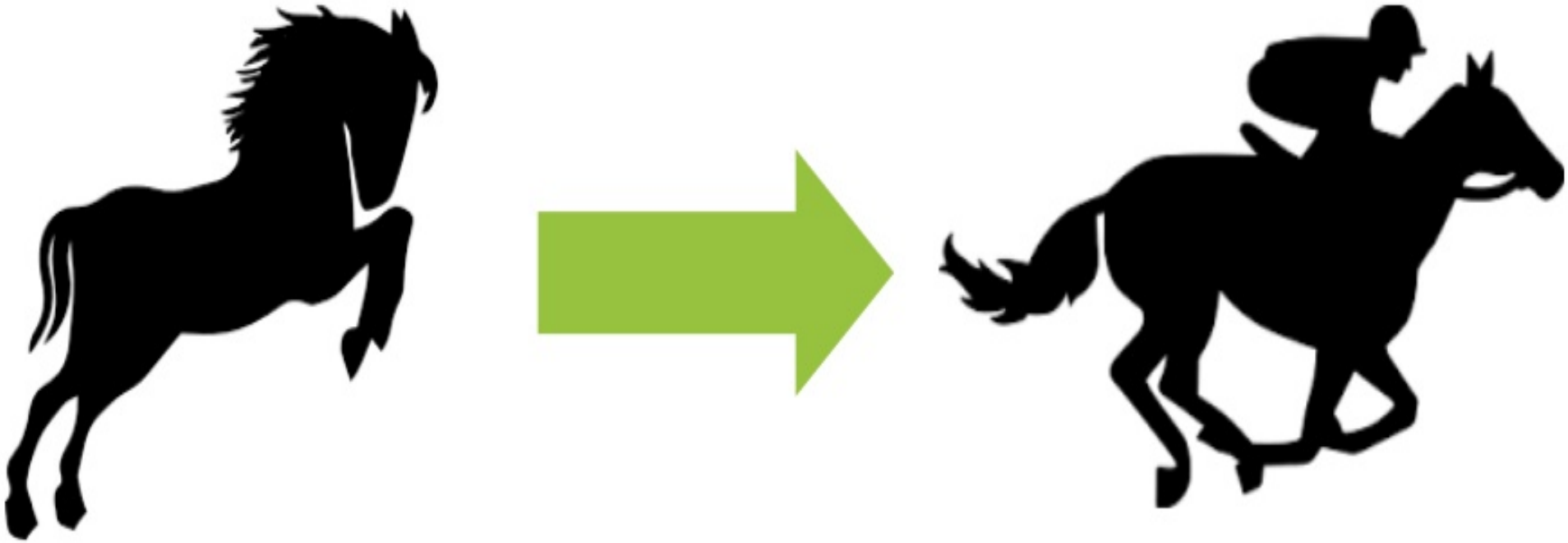
- Analytics has been the backbone of the business.

- We perform the full spectrum of analytics
  - What happened (Descriptive)
  - Why it happened (Diagnostic)
  - What will happen (Predictive)
  - What should we do (Prescriptive)

- *Our barn has different breeds of horses that are trained to win different forms of the sport. Be it Show jumping, Endurance racing or Dressage.*

The Scenario

The challenge



Survival instinct

Winning habit

The Challenge - notes

- We have a smart custom analytics solution called Offer Engine
  - Crunches ~20billion rows of transaction data.
  - Calculates probability and ranks 30 billions Customer-Product combinations
  - Proven to produce impressive Marketing RoI for our clients.

- It is an algorithm built using SQL running on in-memory MPP database + shell scripts.
- We want productise & implement across our customer base.
- Major technical hurdles on development. Less flexibility to use different data sources with different velocity. Prolonged time to take code from development to production.

- *In short we wanted to find this new breed of horse which has more speed, agility & flexibility so that it can win this personalisation game anywhere & not just in one field.*

The Objective

- Rewrite the award winning concept

  "Best Use of Customer Analytics/Data in a Loyalty Programme " – 2014 Loyalty awards

- Create a flexible framework

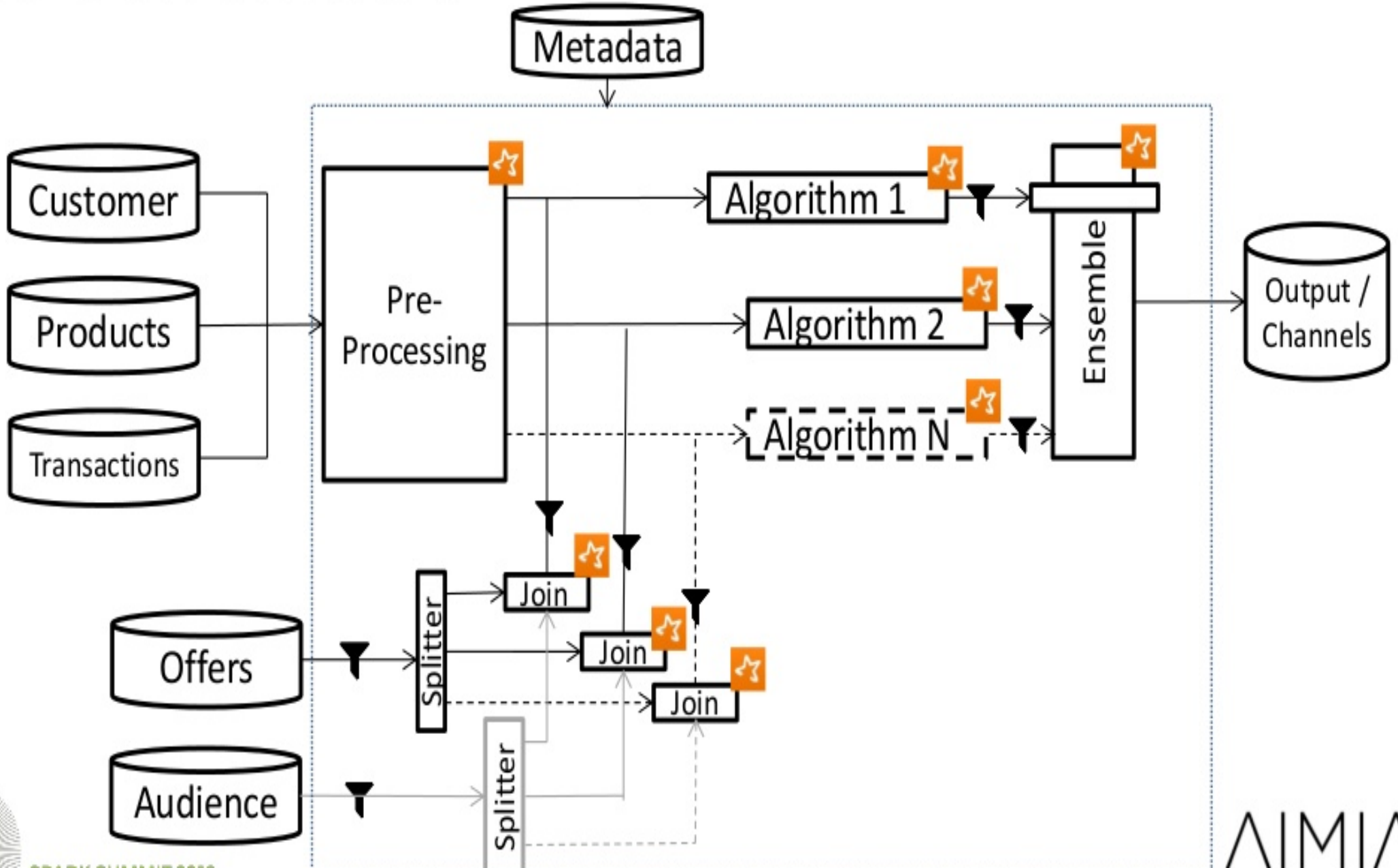
- Custom code to a configurable product

- Better performance at a lower cost

AIMIA

The Journey

The schematic



The Schematic - notes

- The transaction & campaign data from the Enterprise Data Warehouse gets pre-processed to feed to the algorithm.
- The algorithms get trained on the training data from the above.
- When a set of campaign audience & the current offers in store are available, they are sent to different algorithms based on the configuration.
- An Ensemble aggregates & produce the final ranking.

- *It was so obvious to rebuild the key components of the pipeline in Spark to make it a successful product after what we heard in Spark meetups.*
- *The best machine learning implementation using Spark presented in these meet ups from various domains such as banking to e-commerce were convincing.*

The Roles

The Owner  The Sponsor

The Trainer  **The Sparksman**

The Jockey  The field expert

The bettor  The client

The Roles – notes (1)

- The challenge in enterprise adoption of Spark is different. It is a high stakes game.
- Understanding the roles is a key first step to become a great Sparksman.

- The Owner / The Sponsor:
  - Probably the most important person to make your project a success.
  - They are a partnership or a single owner. It is really nice to have a single owner to start with for ease of communication.
  - Need to convince the identified technology is best suited.
  - *You better have the perfect statistics & proof that your new breed of horse is best suited to win even before asking to buy one to test it.*

- The Trainer / The Sparksman
  - We who makes this tech work for our product with thorough knowledge of the market.
  - *The horseman who truly understands the breed & prepares to win the specific sport.*

The Roles – notes (2)

- The Jockey / The field expert:
  - The operations team that makes the product run at scale & deals with issues.
  - The marketing/ sales team that exactly understands the pulse of client.
  - They know the + & - of the product.
  - They need to believe this new tech makes their life easy & better.
  - *Well the Jockey plays a key role in making the horse win the race and it is a perfect partnership between the horseman & jockey is required to make any breed win.*

- The bettor / The client
  - The client collects the market info and purchases the best product & expect it to help us win their customers. They want your product to be flexible, performant & secure.
  - *The bettor analyses the horse info and bet on it for few races. He wants to win big money. They want your horse to be the best of breed, well fit & shows potential.*

The stages

Fear

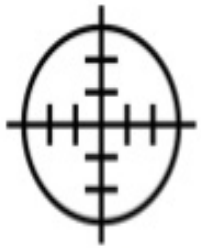
Trust

Comfort

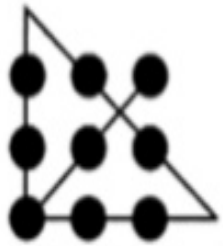
The Stage – notes

- Like it or not, we need to go through the following three stages for success.
- Fear
  - Established technology with expertise available who can bring the best in it. Why change?
  - It is not easy to make the tech work on-prem especially with lack of expertise.
  - *When you train a young horse, it is natural that we will have the fear of understanding what goes on its head and find the best environment to make it shine.*
- Trust
  - This comes with well designed experiments of your different queries.
  - So be prepared for early stage failures when you set up Spark on your laptop, a cluster on VMs on prem without any expertise.
  - Try with leading Spark distribution vendors to make the ramp up easy and painless.
  - *In the right environment with a natural horseman for the horse to gain trust & show its prowess quick.*
- Comfort
  - Need to simulate real scenario & benchmark to get comfortable. Eg. Understand effect of data skewness in full data.
  - *The horse needs few real race experience for it to get used to other horses & the audience.*

The Approach characteristics



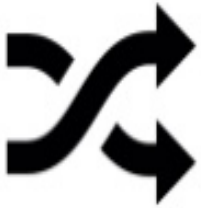
Strong Focus



Unconventional

Patience

Positivity



Timely course correction

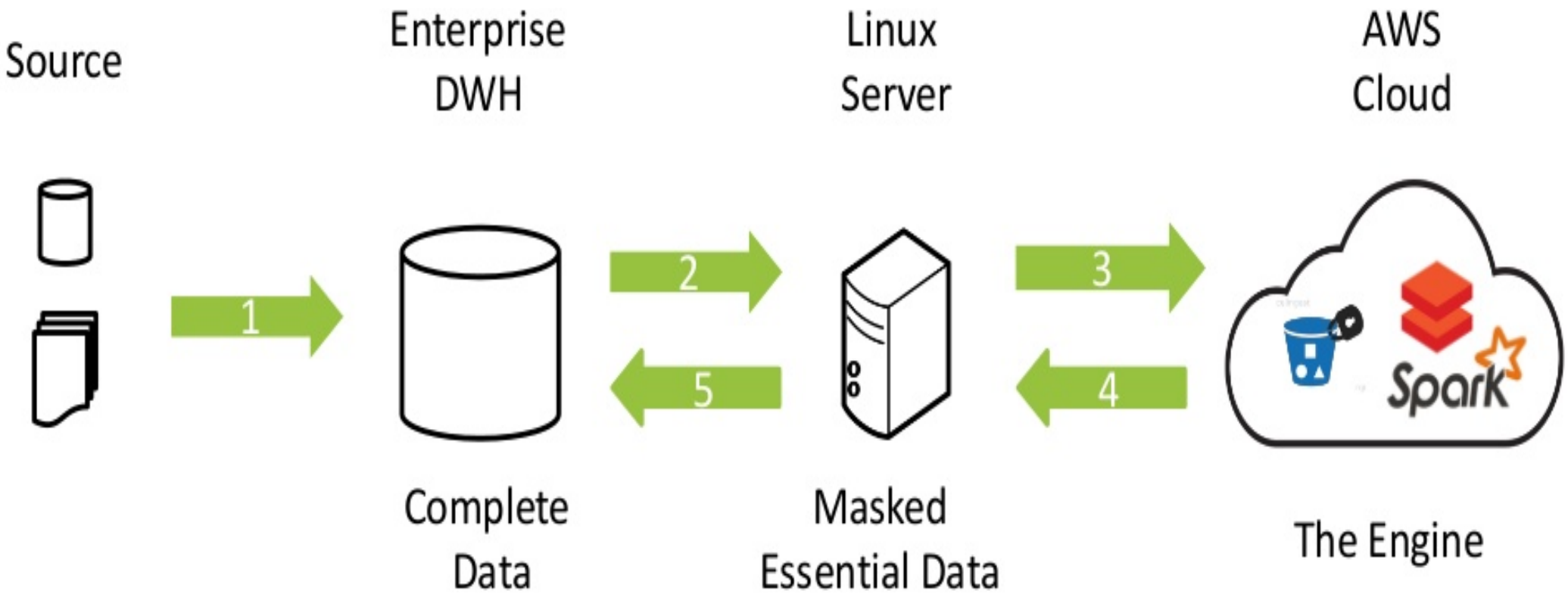
The characteristics – notes (1)

- Here are the top 5 characteristics of a Sparksman
- Strong Focus
  - We all know it is important for any project. But this needs higher emphasis for Spark Projects. Why?
  - Spark as generic execution framework can solve lot of different pieces of the analytics pipeline. (ETL, Data science, Streaming or even other associated benefits such as dynamic & linear scalability).
  - It is easy to get distracted with the capability of Spark. So, define key problem and work to produce solution for it.
  - *You pick the sport you are preparing the horse for and teach the essential skills needed to do it. As the horse has it inherently the nature to run, jump & dance, we should not try to make the same horse trained for race, show jump & dressage all at the same time.*
- Unconventional Approach
  - If you come from a traditional RDBMS/ SQL background to manipulate data, be prepared to unlearn.
  - If you would like to use machine learning concepts, taking the same approach as you took in a SQL engine on MPP database will not work or give the best benefits you expect from Spark.
  - Eg. We used the predicate pushdown concept to make faster & efficient dataframes wherever possible.
  - *While you may not figure out what is happening, if you don't try new options with your horse, you will not know what works & what not. Great trainers have always tried unorthodox things.*

The characteristics – notes (2)

- Patience & Positivity
  - Do not give up when it does not work although it is incredibly boring at time to the same thing again.
  - Repeat the task with tuning one parameter or one aspect of the code at a time. Repetition is key.
  - There are times at which the it won't work for things you don't know or not have control of.
  - Eg. Many a times our code worked for reasonable sized data & scaled linearly but breaks down at some point. GC times were suddenly high.
  - Various approaches for many days did not help until we figured out it is not the code or Spark but data skewness is the issue. So giving the benefit of doubt to the tech worked.
  - *Horses resist to do tasks and sometimes their behaviour bemuses and frustrates you. You cannot give up at the first sign of resistance and need to have the patience. Also great trainers give benefit of doubt to the horse and be positive.*
- Timely course correction
  - Keep the emotion out of the solution. Don't be rigid about your solution.
  - It is important to know when to stop repeating the execution by tweaking parameters & take a new approach to re-write the code.
  - Eg: When our code tweaking didn't work, we figured out some aggregation fails within a window function on the full data set. So, we need to avoid window function & change approach when the dataframe is massive.
  - *Great trainers keep their emotion out & never take frustration on the horse but do some corrections to their approach say by using a stronger bridle or different crop.*

The set-up


Source
Enterprise DWH
Linux Server
AWS Cloud
1
2
3
4
5
Spark
Complete Data
Masked Essential Data
The Engine

The Results

The Achievement

80 - 85% reduction in code base

50% less Memory & 25% less Compute used

20% performance gain (with intermediate stages on disk)

Huge savings on cost per run

The enabler

Language choice

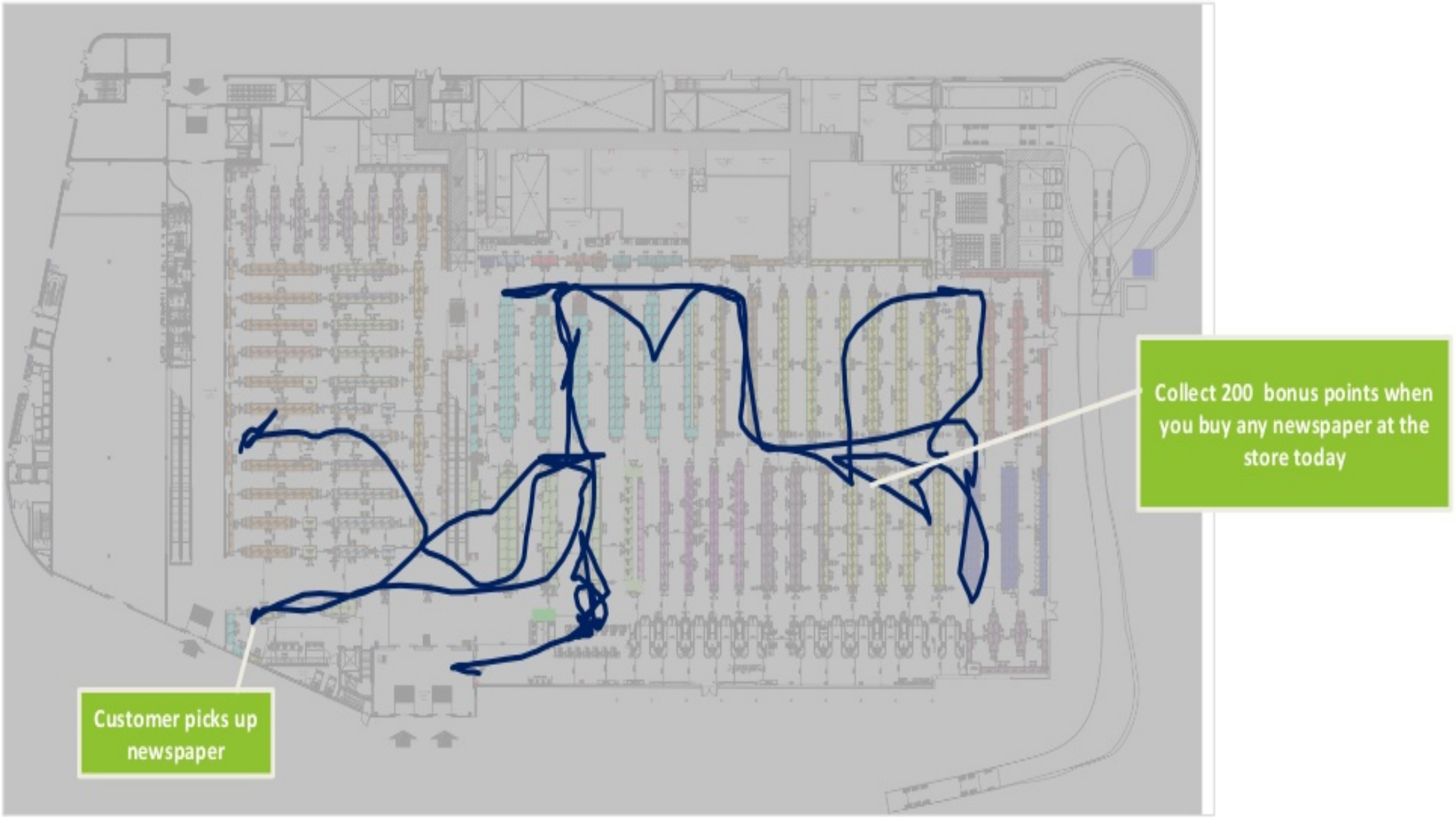
Large scale optimised Spark clusters

Deep Spark Expertise

Dev tool with Integrated Visualisation & Collaboration

Unified web interface for Dev, Test & Operations

ML Lib


Collect 200 bonus points when you buy any newspaper at the store today
Customer picks up newspaper

One key takeaway

"There is no mysticism, no magic, or only one method in the realm of good horsemanship.

It's knowing that everything you think you know about horses may change with the very next horse."

-Buck Brannaman

Final few slides - notes

- We will start using real time data like weather / location of the customer in our personalisation.
- Based on our experiments we know it is a lot of data to consume & analyse in real time.
- This makes Spark very appealing, given 2.0 introduces structured streaming.
- Also the end-to-end security roadmap from databricks might remove our need for data masking.
- Spark projects vary in size, shape & complexity. Spark as a technology is evolving at great speed.
- So, be prepared to unlearn & re-learn, be willing & you can make your analytics enterprise shine with Spark.

AIMIA

Credits

Musa Bilal for the inspiration

Prasad Deshpande for sharing the passion

Stuart Pearson for true Laissez-faire leadership

John Menhinick & Simon Hawkes for unparalleled support

THANK YOU.

@krajeshiyer

www.linkedin.com/in/rajkri

SPARK SUMMIT 2016
DATA SCIENCE AND ENGINEERING AT SCALE