The Hoops Hub: Leveraging Data for Competitive Advantage

The Data Analytics Department

Data Warehouse Specification for The Hoops Hub

Version 1.0

May 6th, 2023

Table of Contents

Introduction

The sports industry, like many others, has always been a very competitive market with the constant need to gain an advantage over competitors. With the rise of big data, the sports industry is altering the way data is collected, stored, and analyzed. Leveraging this data will allow players, teams, and organizations to improve individual performance, the team overall, or make predictions for upcoming talent. This information can also be shared with broadcasters to expand marketing and promote fan engagement.

The goal of this data warehouse is to create a comprehensive and scalable data warehouse solution for the basketball/sports industry that will enable stakeholders to make data-driven decisions and gain insights into performance, fan engagement, and revenue growth. It will leverage the power of big data and drive success in the industry.

Purpose

The purpose of this document is to serve as a project specification document for the design of a data warehouse pertaining to the sports industry. It is to provide guidance for the development team for defining data problems, documenting the process, and requirements of the data warehouse. The content of this document contains the following sections: project summary, requirements definition, considerations, and change log. Within the summary and requirements sections the objectives, scope, references, outstanding issues, goals, and business questions. Lastly, usability, system security, and data requirements are covered.

Project Summary

A. Objectives

1. Improve data quality.
2. Develop business intelligence reports.
3. Design data warehouse that is scalable
4. Structure data models that will enable efficient data retrieval.
5. Ensure the information in the data warehouse is secure.

B. Scope

The scope of this data warehouse design project includes the development of a data warehouse system to address the outstanding issues and requirements of the sports industry.

C. References

D. Bhatnagar and S. Urolagin, "Data Warehousing for Formula One (Racing) Popularity Rating Using Pentaho Tools," 2021 IEEE 6th International Conference on Computing, Communication and Automation (ICCCA), Arad, Romania, 2021, pp. 1-7, doi: 10.1109/ICCCA52192.2021.9666247.

G. Kaur and G. Jagdev, "Analyzing and Exploring the Impact of Big Data Analytics in Sports Science," 2020 Indo – Taiwan 2nd International Conference on Computing, Analytics and Networks (Indo-Taiwan ICAN), Rajpura, India, 2020, pp. 218-224, doi: 10.1109/Indo- TaiwanICAN48429.2020.9181320.

K. Apostolou and C. Tjortjis, "Sports Analytics algorithms for performance prediction," 2019 10th International Conference on Information, Intelligence, Systems and Applications (IISA), Patras, Greece, 2019, pp. 1-4, doi: 10.1109/IISA.2019.8900754.

L. Cai, C. Zhao and X. Wang, "Situation and lessons of application of NBA big data technology," 2021 International Conference on Information Technology and Contemporary Sports (TCS), Guangzhou, China, 2021, pp. 228-231, doi: 10.1109/TCS52929.2021.00054.

Ponniah, P. (2010). Data Warehousing Fundamentals for IT Professionals (Second Edition). Hoboken, NJ: John Wiley & Sons.

V. Bhatt, U. Aggarwal and C. N. S. V. Kumar, "Sports Data Visualization and Betting," 2022 International Conference on Smart Generation Computing, Communication and Networking (SMART GENCON), Bangalore, India, 2022, pp. 1-6, doi: 10.1109/SMARTGENCON56628.2022.10083831.

D. Outstanding Issues

The outstanding issues within the sports industry include data quality, velocity, variety, volume, and value. There are also limitations on data collection depending on the rules and regulations of that specific sports organization.

Data Quality: Sports data can be extracted from various sources including from official sports teams/organizations and official sports data providers. Data can also be extracted from sports fantasy/betting, broadcasting, and other media. However, some media can contain some opinionated and misleading information. Social media is the number one culprit of misinformation. Often times, social media is not used correctly, and sport parody accounts are more common in today's society. With the number of fake websites floating around, the accuracy of data is a challenge within the sports industry.

Volume and Variety: The volume of sheer information has been growing exponentially over the past decade. In addition to the variety of which data can come in, a typical data set can be as high as multiple petabytes. In terms of variety, generated data can include raw, structured, and unstructured data. In terms of data type, information can come in text type, audio, video, and pictorial form. The data warehousing system must be able to navigate between the various variety and the volume of data.

Velocity: In addition to the volume of information increasing, the speed at which it is being generated has also been increasing as technology advances. The design of the data warehouse requires an infrastructure that will keep up with the velocity of data.

Value: Even with a great amount of information it has no use or benefit it has no value. Sports teams and organizations are looking for an advantage over their competition by analyzing player stats, team performances, etc. Data with worth assists in the growth or organization and development of players.

Rules and Regulations: The power of wearable technology can fill many of the gaps when it comes to data that cannot be gathered easily. Wearable technology can access the physical conditions of the players, courts/field/arena, and psychological factors such as referees and fan support. Some professional organizations are worried about the safety of both the player and the sport itself, so they do not allow the use of these devices during the game.

Requirements Definition

A. Goals

1. *Improve data quality:* Data is to be collected from reliable and valid resources, perform data validation and cleaning processes, and perform regular data quality checks.
2. *Deliver business value:* Identify and utilize key performance indicators that align with business goals.
3. *Enable data analysis:* Develop a dashboard to provide insights. Use other reporting tools and technologies like OLAP to enable data exploration.
4. *Enhance data velocity:* Ensure data is captured as close to real-time as possible to support fast data retrieval and support business decision making.
5. *Scalability:* Implement a data warehouse architecture that can support increasing data volume. Develop a data model that can adapt to growth and change over time.

6. *Data security:* Implement access control to prevent unauthorized access and ensure data is backed up regularly to prevent data loss.
7. *Handle volume and variety of data:* Implement a data warehouse architecture that supports large volumes of data in a variety of forms.

B. Usability Requirements

1. *User Interface:* the system must provide a user-friendly interface that will act as a self-service portal and allow users to easily navigate through the data warehouse.
2. *Reporting Tools:* the system will have reporting tools to provide ad hoc reports to improve decision making.
3. *Query Capability:* the system will have capabilities to enable the user with flexibility and customization of queries.
4. *Help Guide:* FAQ and user guide will be provided to assist new users with various functionalities and a tutorial on how software is to be used.

C. System Security Requirements

1. *User authentication:* a login system will be implemented with two factor verification to ensure the user attempting to access the data warehouse is an employee.
2. *Access Control:* role-based access control will be implemented to ensure that users attempting to access data is of clearance.
3. *Data Classification:* highly sensitive data must be classified as such so the system knows what information to protect, and users know what information is the most detrimental if leaked.
4. *Employee Training:* training for each employee is to be provided to ensure employees possess the skills needed for performance. It will also improve performance and reduce the risk of leaked information and/or security breaches.
5. *Intrusion Detection:* security precautions will be installed to prohibit breaches. Security issues and other vulnerabilities in the system will be reported.

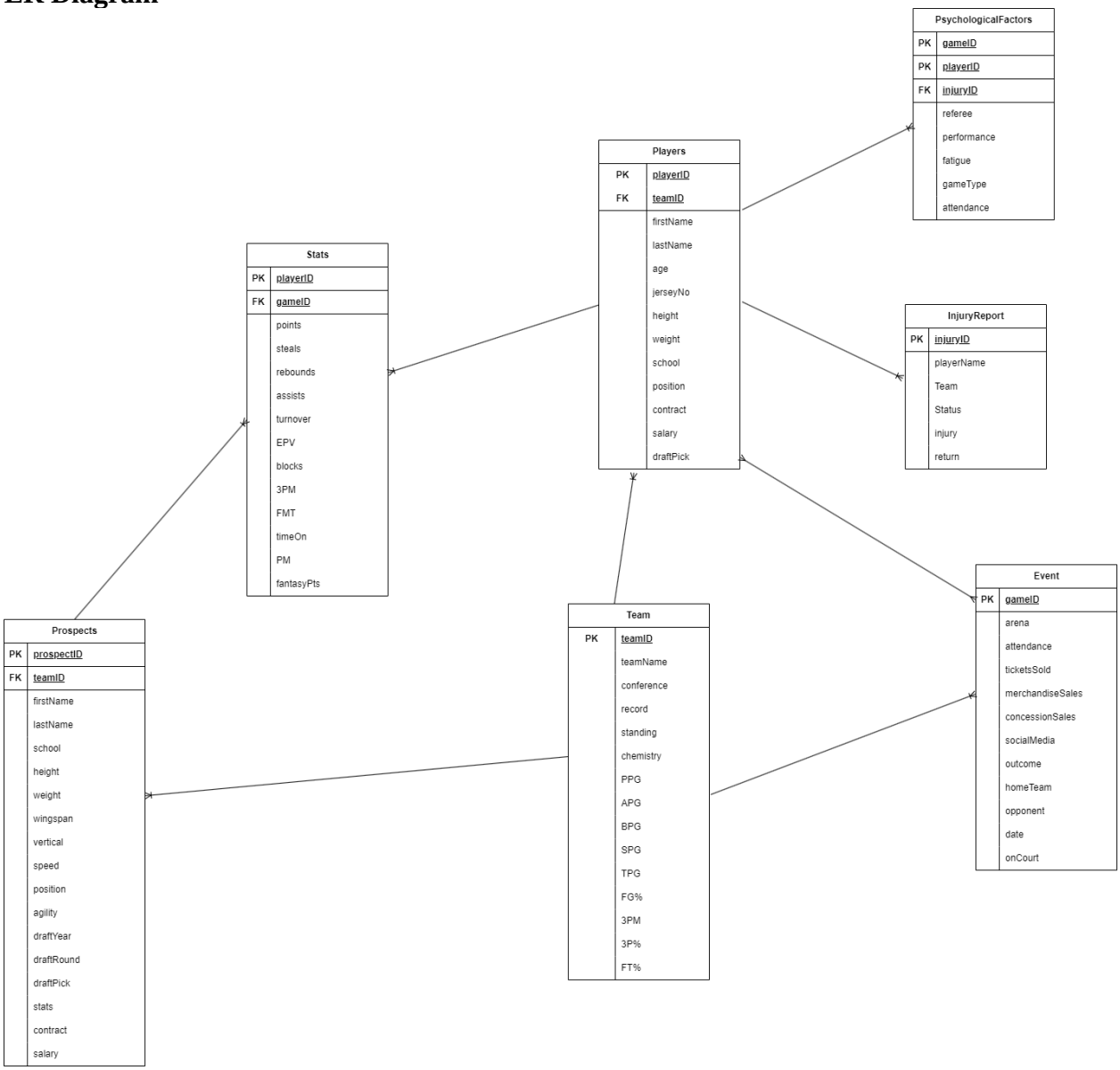
D. Business Questions

1. How are individual players performing in terms of KPI metrics?
2. How are players performances with certain teammates on the floor with them versus without?
3. How is the overall team performance?
4. What factors or trends does the team and/or player have?
5. How is the team and/or player rank in comparison to other teams and/or players in certain categories (i.e. defense, turnovers, scoring, FT made, steals etc.)?
6. What common health issues are faced by players?
7. What psychological factors can affect a player's performance or sway a game's outcome?
8. How is fan engagement? What factors can negatively and/or positively affect fan engagement?
9. What is the expected possession value (EPV) of a player?

10. How can this data be used as a predictive tool for betting agencies, fantasy leagues and prospect (draft) players?

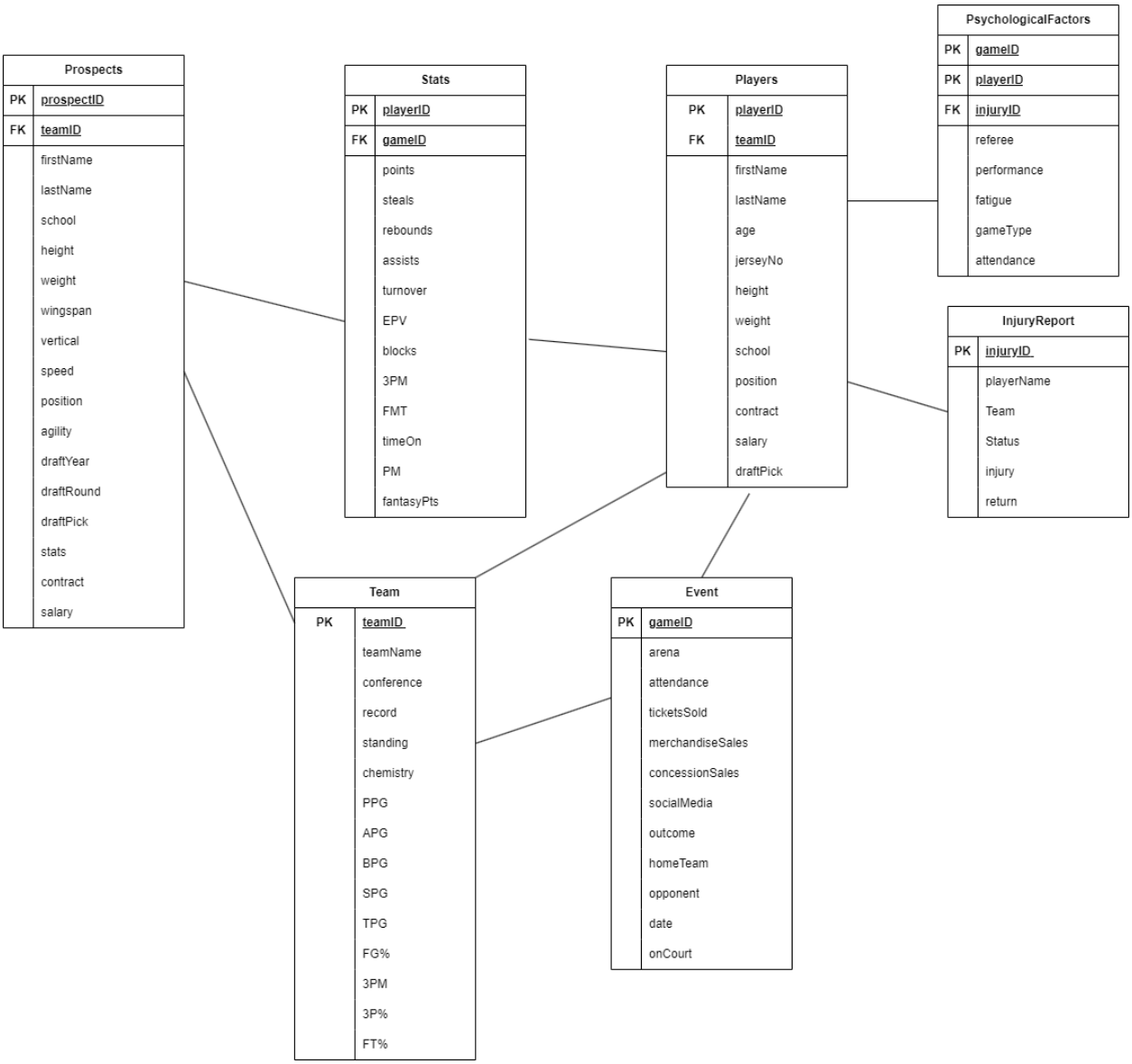
E. Data Requirements

ER Diagram



Database Design

The hoops hub data warehouse is designed as a relational database. It includes a table that can be seen below and the lines in the diagram signify the relationships between the tables. Primary and foreign keys are utilized to help users easily retrieve related information.



Data Dictionaries

Table Name: Players

Field Name	Data Type	Constraint	Description	Example
playerID	Int	Primary key	The unique identifier for a player	P127009
firstName	Text		The players first name	Lebron
lastName	Text		The players last name	James

teamID	Text	Foreign key	The unique identifier for the team the player plays on	T015
age	Int		The age of the player	38
jerseyNo	Int		The players jersey number	6
Height	Text		The players height	6’9”
Weight	Int		The players weight	250
School	Text		The school the player got drafter from	St. Vincent-St. Mary Highschool
Position	Text		The position the player plays	SF
contract	Text		the details of the players contract including length, trade clauses and other perks or incentives	5 years
Salary	Float		The amount of money the player is being paid via their contract	44,000,000
draftPick	Text		The year, round and pick the player was drafted	2003: Rd1, Pk1

Table Name: Stats

Field Name	Data Type	Constraint	Description	Example
playerID	int	Primary key	the unique identifier for the player	P127009
gameID	Int	Foreign key	the unique identifier for the game the player played in	E005111
Points	Float		the running average number of points the player scores	28.5
Steals	Float		the running average of steals a player gets per game	1.2
Rebounds	Float		the running average of rebounds a player gets per game	8.3
Assists	Float		The running average of assists a player gets per game	6.8
Turnover	Float		The running average of turnovers a player commits per game	5.6
EPV	Float	Optional	The expected possession value by the player	.88
Blocks	Float		The running average of blocks the player gets per game	4.7
3PM	Int		The running total number of three pointers made by the player	2,243

FTM	Int		The running total of free throws made by the player	8,542
timeOn	Int		The running total number of minutes played	38
PM	Int		plus/minus stat (+/-) to measure players impact on the courts versus off the court	+10
FantasyPts	Int	optional	The running total of fantasy points accumulated by the player	5

Table Name: Team

Field Name	Data Type	Constraint	Description	Example
teamID	Int	Primary Key	The unique identifier for the team	T015
teamName	Text		The name of the team	Los Angeles Lakers
Conference	Text		The name of the conference the team plays in	Western
Record	Text		A running record off wins/losses	52-28
Standing	Int		The teams rank	7
Chemistry	Float		The	78.9
PPG	Float		The average points per game made by the team	98.5
RPG	Float		The average rebounds per game made by the team	50.5
APG	Float		The average assists per game made by the team	20.3
BPG	Float		The average blocks per game made by the team	9.8
SPG	Float		The average steals per game made by the team	3.2
TPG	Float		The average turnovers committed per game by the team	8.9
FG%	Float		The field goal percentage of the team	46.7
3PM	Int		The number of three pointers made by the team	6
3P%	Float		The three-point field goal percentage made by the team	24.0
FT%	Float		The free throw percentage made by the team	86.2

Table Name: InjuryReport

Field Name	Data Type	Constraint	Description	Example
injuryID	Int	Primary key	The unique identifier for the injury report	I730981
playerName	Text		The name of the player who has an injury	Jimmy Butler
teamID	Int	Foreign key	The unique identifier of the team the player plays on	T43
Status	Text		The players current status can be one of the following: available, possible, doubtful, out	Out
Injury	Text		A description of the body part and type of injury the player is faced with	Ankle (sprined)
Return	Text	Optional	An estimated time the player will return to action	3 weeks

Table Name: Event

Field Name	Data Type	Constraint	Description	Example
gameID	int	Primary key	The unique identifier for the game the game the player played in	E005111
arena	Text		The name of the arena the game was played in	Spectrum Center
attendance	Int		The total number of fans in attendance for the game	18,938
ticketSold	Int		The total number of tickets sold	19,012
merchandiseSales	Float		The total number of sales made on merchandise	1,724.00
concessionSales	Float		The total number of sales made at concessions	5,874.10
socialMedia	Int		The number of like, comments, shares, posts pertaining to the current event	300,530
outcome	Text		Records either a win or a loss based on the home team	L

homeTeam	Text		Name of the home team for the game	Charlotte Hornets
opponent	Text		Name of the away team for the game	Memphis Grizzlies
date	Date		The date the game takes place	3/15/2023
onCourt	Text		The list of players on the court (in the game) at a given time	Lamelo Ball, Kelly Oubre, Gordon Hayward, Terry Roizer, PJ Washington

Table Name: Prospects

Field Name	Data Type	Constraint	Description	Example
prospectID	Int	Primary key	Unique identifier for the given prospect	PR827301
teamID	Int	Foreign key	The unique identifier for the team the prospect will be drafted by	T13
firstName	Text		The first name of the prospect	Jake
lastName	Text		The last name of the prospect	Maverick
School	Text		The name of the school the prospect will be drafted from	Ben Smith Highschool
Height	Text		The height of the prospect	6’5”
weight	Int		The weight of the prospect	195
wingspan	Text		The wingspan in inches of the prospect	23.67
vertical	Text		The vertical in inches of the prospect	36.05
speed	Float		The speed of the prospect conducting a ¾ court sprint	2.85
position	Text		The position the prospect plays	SG
agility	Text		A description of the prospect’s agility either can be below, average, or above	average
draftYear	Int		The year in which the prospect was drafted	2023
draftRound	Int		The round number the prospect was drafted	1
draftPick	Int		The pick number the	25

			prospect was drafted	
contract	text		the details of the players contract including length, trade clauses and other perks or incentives	5 year , no trade clauses
salary	Float		The amount of money the player is being paid via their contract	950,000.00

Table Name: PsychologicalFactors

Field Name	Data Type	Constraint	Description	Example
playerID	Int	Primary key	The unique identifier for the given player	P086314
gameID	Int	Primary key	The unique identifier for the event/game	E001116
injuryID	Int	Foreign key	The unique identifier for the injury report on a given player	I632980
referee	Text		The name of the referee scheduled for the game	Scott Foster
performance	Text		Rating on the current performance of the player. Can either be excellent, good, fair, or poor.	good
Fatigue	Float		The fatigue percentage on the given player	34.00
gameType	Text		The type of game either home or away	away
attendance	Int		The number of people in attendance for the game	19,348

F. Design Constraints

1. Budget Limitations:
   - Identify areas in data analytics department where additional resources may be needed.
   - Set aside budget for hiring people with given talent/skill set.
2. Organizational Policies:
   - Some data on prospects will be missing due to optical tracking not being available at that level.
   - Organizations must all use SportVU for tracking data for consistency measures
3. Resource Limitations:
   - Use efficient hardware and software components needed for data warehouse.

4. Time Constraint:
    - Prioritize efficient and timely data loading process to ensure the availability of up to date information for real time data analysis and reporting.

Considerations

- Teams in sports industry has difficulty reserving funds for data scientists who help draw conclusions and make predictions.
- Optical tracking data is currently only available in soccer and basketball. SportVU is only available in NBA. Other organizations do not use wearable technology and/or limit the use of tracking data.
- Typical data sets can be as high as multiple petabytes. Storage system requirements must be in consideration in the design phase.
- ESPN reporters, analysts, and broadcasters use this information to provide commentary, analysis, and updates before, during, and after live game events. Queries must be optimized in order to enhance performance and provide data in real time.

Appendix

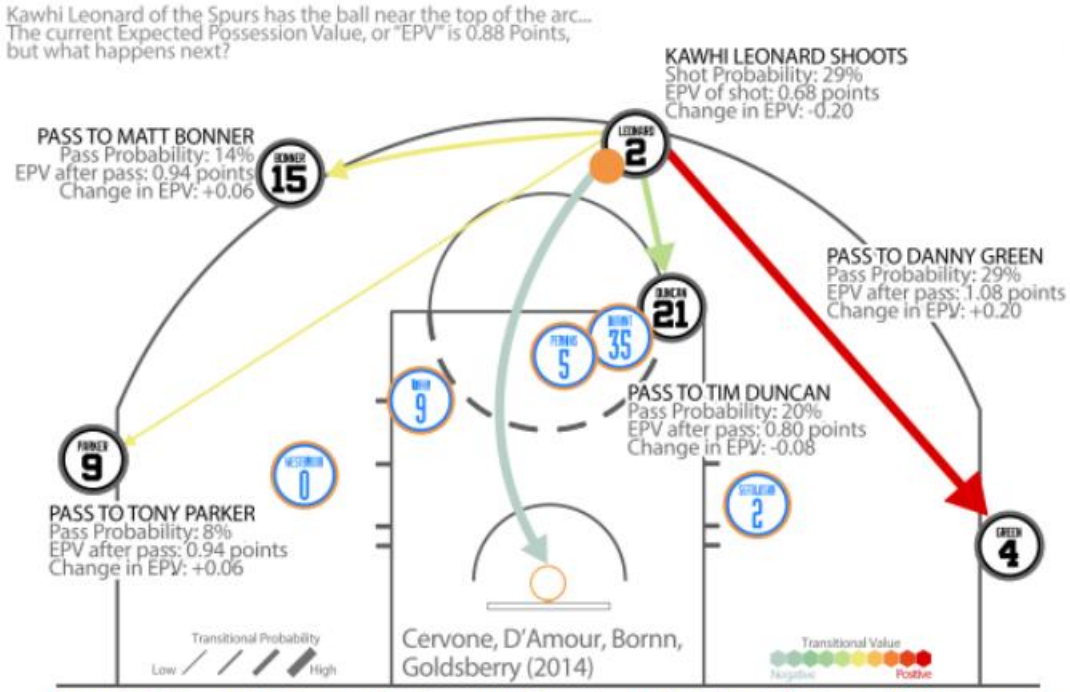


Figure 1. Diagram of EPV as a weighted average of the values of the ballcarrier's (Leonard's) decisions and the probability of making each decision. We also consider the possibility of Leonard dribbling to a different area or driving toward the basket, as well as turning the ball over, but these are omitted from the above diagram for conceptual clarity.

Document Change Log

Change Date	Version	CR #	Change Description	Author and Organization
02/03/23	1.0		Initial creation.	Taylor Headen
02/10/23	1.1		Introduction and Scope completion	Taylor Headen
2/24/23	1.2		Resources added	Taylor Headen
3/4/23	1.3		Project summary creation	Taylor Headen
3/8/23	1.3.1		Objective, Scope and Outstanding issues completion	Taylor Headen
3/17	1.4		Requirement definition creation	Taylor Headen
3/30/23	1.4.1		Goals and business question completion	Taylor Headen

4/10/23	1.4.2		Usability and security requirements completion	Taylor Headen
4/21/23	1.4.3		Data dictionary completion	Taylor Headen
4/27/23	1.4.4		Requirements completion	Taylor Headen
4/29/23	1.5		Appendix added, consideration completion	Taylor Headen
5/4/23	1.6		Formatting and Finishing touches	Taylor Headen

Data Warehouse Design Template (n.d.). Retrieved from https://www.google.com/search?client=safari&rls=en&ei=h2HeWv6fOY_-zgK5vrOIDQ&q=data+warehouse+design+document+template+&oq=data+warehouse+design+document+template+&gs_l=psy-ab.3..0i30k1j0i8i30k1.4517.15588.0.15915.16.12.4.0.0.0.129.1130.9j3.12.0....0...1.1.64.psy-ab..0.13.1029...0i22i30k1j0i7i30k1j0i8i7i30k1j0i13i30k1j0i8i13i30k1.0.JMhxh9fWO5g